Some Tools for Robustifying Econometric Analyses

Victor Hoornweg & Philip Hans Franses

November 2013

Abstract

We use automated algorithms to update and evaluate *ad hoc* judgments that are made in applied econometrics. Such an application of automated algorithms robustifies empirical econometric analyses, it achieves lower and more consistent prediction errors, and it helps to prevent data snooping. Tools are introduced to evaluate the algorithm, to see how configurations are updated by the algorithm, to study how forecasting accuracy is affected by the choice of configurations, and to find out which configurations can safely be ignored in order to increase the speed of the algorithm. In our case study we develop an algorithm that updates *ad hoc* judgments that are made in Cápistran and Timmermann's (2009) attempt to beat the mean survey forecast. Many of these *ad hoc* judgments are often made in time series forecasting and have hitherto been overlooked. We show that our algorithm improves their models and at the same time we further robustify the stylized fact that the mean survey forecast is indeed difficult to beat.

ECONOMETRIC INSTITUTE REPORT

EI 2013-33

1 Introduction

Empirical econometric analysis can concern case studies or stylized facts. Examples of case studies are the provision of forecasts for certain macroeconomic variables for a particular country and a particular time frame, or an examination of the potential drivers of recessions using leading indicators. Examples of stylized facts are that economic time series have stochastic trends rather than deterministic trends and that the average of a range of forecasts seems to be hard to beat in terms of accuracy. In both situations the analyst usually aims to draw conclusions that are robust to changes in the premises. For example, macroeconomic forecasts for the next year should preferably be independent from the starting point of the sample used to generate these forecasts. And, the general finding of stochastic trends should best be independent from the type of test used. In general, one would thus want that empirical econometric analysis is independent from the various statistical decisions that analysts have to make, at least, as much as possible. In the present paper we aim to meet this desire by proposing a set of tools involving automated algorithms that robustifies the analysis and indicates where path-dependency is strongest.

There is one area where automated algorithms are already quite popular and this concerns the selection of variables in a regression model. An example of such an automated algorithm is the general-to specific (Gets) procedure (Hendry and Krolzig, 2005), which starts from a large model and subsequently deletes the worst variables (for instance in terms of p-values) until some criterion is met. Leamer (1978) focuses on the selection of variables in a Bayesian setting. In case there is uncertainty about which statistical method is appropriate, like a vector autoregression (VAR) or principal components analysis (PCA), automated algorithms have been applied to select and combine the best forecasts of each model, see Bache et al. (2012). Automated algorithms have also been used to optimize over the estimation sample, by averaging over forecasts generated

by different starting points of the treatment sample, see Hendry (2006) and Pesaran and Timmermann (2007).

Aside from these well appreciated applications of automated algorithms, many of the most basic statistical decisions are not yet updated by data with the help of an automated algorithm. Decisions which are not updated by data in a real-time setting will be called '*ad hoc*'. For instance, when an analyst has to decide upon a parameter like a shrinkage factor in a shrinkage model, the output is typically generated and presented for a few pre-specified parameter choices. A shrinkage factor shrinks a combined forecast with individual weights to a combined forecast with equal weights. Stock and Watson (2004) for example consider three shrinkage rates and conclude that "the shrinkage forecasts are not robust: for some countries and horizons they perform well, but for others they perform quite poorly" (*ibid*, pp. 514). We tend to believe that an automated algorithm could have been useful here to optimize over the shrinkage level in a real-time setting and to set the scenes for detailed investigations about the extent to which shrinkage rates influence forecasting accuracy.[1]

As another example where automated algorithms could be useful, think of the weights that analysts have to assign to forecasts when combining h-period-ahead forecasts. Usually, analysts only consider the h-period-ahead prediction errors, thereby neglecting that $(h+1)$-period-ahead prediction errors might also be informative about a model's forecasting ability, because further ahead forecasts are often more difficult to create. At the same time, $(h-1)$-period-ahead forecast errors could be indicative of a model's *current* forecasting performance as these forecasts precede the h-step-ahead forecasts. One might thus wish to implement an automated algorithm to determine which h^*-step-ahead forecasts are relevant for assigning weights to individual models. In a similar vein, the combination weight of a model can also be based on the model's prediction errors of other variables than the dependent variable, particularly when little

[1] In a similar vein, a grid search is often used to find the smoothing parameter for exponential smoothing in Gardner (1985, pp. 7).

data is available.

When an algorithm is used to select the optimal models, an *ad hoc* decision is oftentimes made on the number of top-ranked models to be pooled. Typically, analysts either select the single best model, or a weighted average of all available models (see for example Stock and Watson, 2004). Similarly, in seeking the optimal starting point of the treatment sample when there is no estimate of a break date, Pesaran and Timmermann (2007) only consider taking the single best starting point or a weighted average of all of the available starting points. Based on a simulation study, they find that it depends on the timing and nature of the breaks which approach works best. The optimal number of top ranked models or starting points to pool might lie between one and all and might depend on the situation. An automated algorithm can dynamically update the number of top ranked models and starting points to be combined. Further tools could be employed to examine closely how forecasting errors are related to the choice of model or the choice of starting point.

The general strategy to update such *ad hoc* decisions using automated algorithms is easy to execute. For the sake of clarity, we say that a decision occurs when a 'configurations' of a particular 'item' is selected, like setting the starting point (item) of a treatment sample at 1970Q1 (configuration) or the shrinkage rate (item) to 0.5 (configuration). To update a decision about a particular item by data, the analyst first defines a set of candidate configurations for that item, like the shrinkage rate $\phi \in \{0, 0.1, \ldots, 1\}$.[2] Then, pseudo-out-of-sample forecasts are produced for each configuration, resulting in scores that are based on the associated pseudo-out-of sample forecast errors. Finally, the best configuration is selected for producing the out-of-sample forecast. Whenever feasible, we will rank configurations based on their pseudo out of sample performance and subsequently select the optimal number of top-ranked configurations to be

[2]Reviewers can request that a researchers expands the set of candidate configurations to further robustify the analysis and to deal with potential data-snooping behavior in the selection of candidate configurations.

pooled. Note that when one would be confronted with cross-sectional data instead of time series data, one can partition the data in a treatment sample and a validation sample. Furthermore, we would like to point out that an analyst might introduce a more informative prior by taking a weighted average between the score based on pseudo out of sample prediction errors and a score that is pre-specified by the analyst. Indeed, an *ad hoc* decision can be regarded as a really strong prior.

The benefits of employing automated algorithms to update *ad hoc* decisions make are threefold. First, the application of an automated algorithm robustifies empirical econometric analyses and lays bare the uncertainty involved in judgements that are made in real time. Second, by updating the statistician's judgments by data, lower and more consistent prediction errors can be achieved. Third, the application of automated algorithms can lessen data-snooping, whereby a researcher chooses *a posteriori* what the best set of configurations are, such as the choice between a rolling window and an expanding window (Pesaran and Timmermann, 2005).

There are also two main challenges in using automated algorithms. One difficulty is that automated algorithms can turn into a black box. The analyst may have no clue which specifications were selected by the algorithm for each prediction, how much uncertainty was involved in the selection of configurations, and the analyst might not have learned which configurations lead to good forecasting accuracy and which configurations should best be avoided. This information is important, because it helps the analyst to define an informative prior and because it could motivate her to investigate more configurations of certain items or new procedures for selecting configurations. A second difficulty is that the analyst has to curtail the number of candidate configurations for each item. When too many combinations between configurations of different items are considered, the total number of perturbations could become too large, making the algorithm too slow.

It is our our aim to advocate the use of automated algorithms to update *ad hoc* decisions, and we will therefore provide tools for dealing with these two challenges. To shed some light on the otherwise opaque inner workings of the algorithm, we use distribution images to assess how the analyst's decisions are updated by the data. A 'distribution image' shows the weights that are assigned to different configurations for a particular decision at a particular time. It is called after the Matlab command 'image'. An 'accuracy image' will be used to compare the forecasting accuracy of different combinations of configurations. This tells us how *ad hoc* choices affect forecasting accuracy and how the algorithm can be improved. Another plot will be presented which shows how successful the algorithm was in selecting configurations. To tackle the second problem of limiting the number of configurations per item, the analyst needs to know how much predictions vary across different configurations of a particular item. If the analyst defined a set of, say, eleven different shrinkage rates, but the choice of shrinkage rate barely affects final forecasts, then she might reduce the number of candidate shrinkage rates considerably. An 'S-image' is used to show the average amount of variation in predictions across different configurations of an item (S) conditional on configurations of other items.

To illustrate how automated algorithms can be used to provide robust conclusions by updating *ad hoc* decisions, we take the important study of Capistrán and Timmermann (2009) as the running example. Using the USA Survey of Professional Forecasters (SPF), Capistrán and Timmermann evaluate various methods which aim to optimally combine expert forecasts into a single forecast. They conclude that the simple equal-weighted mean is extraordinarily difficult to beat, and we shall take this conclusion as an example of a stylized fact, also as it has been frequently documented in the literature. By updating their *ad hoc* decisions we show that their basic strategies can be improved in terms of forecasting accuracy. In doing so, we will obtain even stronger corroborat-

ing support for the finding of these authors that the equal-weighted forecast is difficult to beat in terms of forecast accuracy.

Case-specific for the Capistrán and Timmermann (2009) study, we document optimal ways to: (a) determine the number of best-ranked forecasting experts [models] that receive a non-zero weight, (b) to strike a balance between individual weighting and equal weighting by detecting the optimal shrinkage factor in a shrinkage model, (c) to identify the relevant start of the treatment sample, (d) to assess whether forecasting errors of other variables are helpful when assigning combination weights to forecasts of one particular variable, (e) to select the most relevant forecast horizons for which the associated forecast errors are to be used for computing forecasts' combination weights, and finally (f) to choose the proper evaluation function. We would like to note that 'optimality' is used relative to an explicit loss function and to the information that is available in real time. To ensure reproducibility, we put a step-by-step explanation of the code in the appendix (A) to this paper.[3]

To be clear, we do not mean to imply that Capistrán and Timmermann displayed data-snooping behavior because we can identify their *ad hoc* choices. By contrast, it is because we appreciate their effort to arrive at a robust stylized fact, that we want to build on it. Moreover, the examples above illustrate that the *ad hoc* choices of Capistrán and Timmermann are decisions that are *generally* made in such a particular way in empirical econometrics.

All in all, there are three main contributions of this paper. We show how updating *ad hoc* decisions using an automated algorithms can lead to more robust, optimal, and honest forecasts. Moreover, many of the *ad hoc* judgments we update in the this paper have, to the best of our knowledge, hitherto been overlooked in empirical econometrics. Furthermore, new tools are presented for analyzing how *ad hoc* decisions were updated by the algorithm and the data.

Our paper is organized as follows. In Section 2 we discuss the various

[3]The code can be downloaded from [to be announced].

relevant aspects of Capistrán and Timmermann (2009). Section 3 introduces the components of our automated algorithm. In Section 4 we present tools that can be used to analyze the automated algorithm in a practical setting. In Section 5 we return to the empirical study and we present our results. Section 6 concludes with suggestions for further research after a critical appraisal of the findings in the present study.

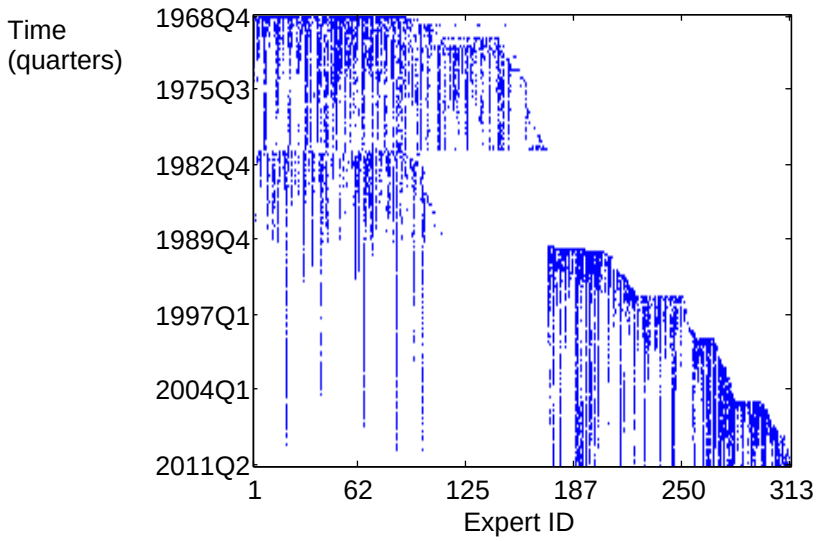
2 Capistrán and Timmermann

The SPF collects predictions of professional forecasters on various key macroeconomic variables. The strength of surveys like the SPF is that its members have diverse backgrounds and that they employ different forecasting techniques. Some of the professionals work in banks or insurance companies, while others are affiliated to a university or a forecasting firm. Some forecasters use leading indicators, econometric models, or an informal approach, while others rely on personal judgments (Zarnowitz and Braun, 1992, pp. 16).

A drawback of these surveys is the frequent exit and entry of individual forecasters, which is illustrated by Figure 1. A blue dot indicates that the expert submitted a forecast in a given quarter. Usually, when several forecasts are available, analysts (econometricians) attempt to optimally combine these forecasts. Note that due to the discontinuity of individual responses in survey data, the often-considered least squares approach to combine forecasts has become infeasible. Because of this, most analysts simply use an equally weighted mean forecast or weights that rely on pseudo mean squared forecast error (pMSFE) (Capistrán and Timmermann, 2009).

Capistrán and Timmermann (2009) (CT) test various ways to improve the mean SPF forecast. They use an Expectation Maximization (EM) algorithm to backfill past missing observations in order to combine forecasts in the extended panel. They attempt to trim forecasts from participants who did not report

Figure 1: Individual responses of USA SPF experts over time



A blue dot indicates that an expert has submitted a forecast at a particular time.

a minimum number of forecasts, under the assumption that poor forecasters would quit earlier. Among these and many more strategies, CT find that only a method whereby a Schwarz Information Criterion (SIC) is used to decide between a bias-adjusted model and the mean survey forecast occasionally improves mean survey forecasts. The authors conclude that, 'in common with empirical findings in the literature, the simple equal-weighted forecast turns out to be extraordinarily difficult to beat' (*ibid.*, pp. 438).

Additional to the mean survey forecast and their SIC bias-adjusted model, we shall present their previous-best forecast, their inverse pMSFE method, and their shrinkage model. Subsequently, we will identify *ad hoc* decisions that are made in these approaches. Finally, it will be discussed how the interaction of these statistical judgments might affect forecasting errors.

The Mean SPF, that is, the mean survey forecast, takes the equal weighted mean of the expert forecasts, written as

$$\bar{Y}_{t+h|t} = \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{Y}^i_{t+h|t}, \tag{1}$$

where $\hat{Y}^i_{t+h|t}$ denotes the h-period-ahead forecast made by forecaster $i = 1, \ldots, N_t$ at time t for some variable Y.

The SIC bias-adjusted model is an example of an automated algorithm whereby the choice between Mean SPF and the bias-adjusted model is updated using a SIC. The bias-adjusted model is given by

$$\tilde{Y}_{t+h|t} = \alpha + \beta \, \bar{Y}_{t+h|t}. \tag{2}$$

The acronym of this model is 'SIC-bias'. The information criterion is calculated as

$$\text{SIC}(p) = \log(s_p^2) + \frac{p \log(v_t)}{v_t},$$

where v_t denotes the size of the evaluation window and where s_p^2 is the maximum likelihood (ML) estimator of the error variance in the model with p regressors under the assumption that the errors are identically and independently distributed according to a normal distribution (Heij et al., 2004, pp. 279).

The previous-best-forecast does exactly what its name suggests, that is, it sets $\hat{Y}^*_{t+h|t} = \hat{Y}^{i^*_t}_{t+h|t}$. The best individual i^*_t is found using a pMSFE criterion, defined by

$$i^*_t = \min_{i=1,\ldots,N_t} \frac{1}{v} \sum_{\tau=t-v+1}^{t} e^2_{\tau,\tau-h,i}, \tag{3}$$

where $e_{\tau,\tau-h,i} = (\hat{Y}^i_{\tau|\tau-h} - Y_\tau)$ is the h-step-ahead forecast error made by forecaster i at time t.

The inverse pMSFE method assigns weights to forecasters who have a sufficiently long track record by taking the inverse of the individual's historical pMSFE, that is,

$$w^i_{h,t} = \frac{(\frac{1}{v}\sum_{\tau=t-v+1}^{t} e^2_{\tau,\tau-h,i})^{-1}}{\sum_{j=1}^{N_t}(\frac{1}{v}\sum_{\tau=t-v+1}^{t} e^2_{\tau,\tau-h,j})^{-1}}. \tag{4}$$

Forecasters are required to have submitted a minimum of ten contiguous forecasts. Weights are estimated for the largest common sample. Forecasters whose

track records are too short receive equal weights. The weights are normalized to sum to 1.

Following Stock and Watson (2004), CT also consider shrinking the estimated weights to the arithmetic average of forecasts, in order to reduce the risk of giving improper weights. Their shrinkage model is applied to least squares estimates of the combination weights. Call $\hat{\omega}_t^i$ the estimated weight of the i^{th} forecaster [model], then the combination weights become

$$\begin{aligned} \omega_t^i &= \xi_t \hat{\omega}_t^i + (1 - \xi_t)\frac{1}{N_t} \\ \xi_t &= \max\left(0, 1 - \kappa \frac{N_t}{v_t - 1 - N_t - 1}\right). \end{aligned} \tag{5}$$

The analyst has to decide on the value of κ, which is a constant that controls the degree of shrinkage towards equal weighting. CT consider $\kappa = 1$ and $\kappa = 0.25$. A large value of κ lowers ξ_t and thereby increases the shrinkage towards equal weights. Furthermore, a sample size v_t that is large relative to the number of forecasts N_t results in less shrinkage towards the mean. As Stock and Watson (2004) note, the shrinkage forecast can be interpreted as a Bayes estimator, applying the principle of indifference (equal-weighting) as a prior.

CT use an expanding window setup whereby the starting point of the treatment sample is in 1981Q3 and the initial size of the treatment sample is thirty observations. The starting point of 1981Q3 instead of 1968Q4 is chosen because some of the variables considered by CT were forecasted by SPF members only after 1981Q2. In their study, the h-quarters-ahead forecasts of no less than fourteen variables are analyzed, where $h = 1, 2, 3, 4$. Unless the initial variable is measured in percentage change, the variables are transformed into growth rates for quarter-to-quarter change, expressed in annualized percentage points as in

$$x_{t+h} = 400 * \ln \frac{X_{t+h}}{X_{t+h-1}}, \tag{6}$$

in order to deal with redefinitions of variables, like changes in base years. The

flash realization of a given quarter, which is published in the subsequent quarter, is used to evaluate the forecasts. This is fair because the forecasters too could only avail themselves of flash realizations. Out of the fourteen variables CT looked at, we shall analyze the Price index of Gross Domestic Product (PGDP) and Nominal Gross Domestic Product (NGDP). For PGDP and NGDP there is data available that starts in 1968Q4, and we will therefore use the full dataset. PGDP is our main variable of interest, because it is one of the few variables whereby the SIC-bias method and the inverse pMSFE method outperformed Mean SPF. In other words, for PGDP, the conclusion that Mean SPF is hard to beat is challenged the most. We want to analyze how *ad hoc* decisions influenced this result and whether it helps to optimize over these choices using an automated algorithm.

Before we turn to discuss the *ad hoc* decisions, we would shortly like to summarize Pesaran and Timmermann's (PT) article on the selection of starting points of the treatment sample (2007). To optimize over the starting point of the treatment sample when there is no estimate of a break date available, they choose to either select the single best starting point, or to combine all of the available starting points based on pMSFE weights. As PT remark, the choice of window size used for evaluating starting points is also important. A large window excludes recent starting points, while a small window gives unreliable estimates. In their simulation study, a window of ten or twenty pseudo out of sample forecasts are used to compute the pMSFE scores of each starting point. When we refer to 'window' in this paper, we mean the window used for evaluating different starting points. Among their benchmark models are the traditional approach of using the first available starting point and another approach that takes the equal-weighted mean of all starting points. Based on a simulation study, they find that the optimal method for selecting starting points depends on the timing, frequency, and nature of the breaks. They also show that it can be optimal to use pre-break data to estimate forecasting models on

data samples subject to structural breaks. Some of the *ad hoc* choices in PT will also be identified below. This allows us to study in more detail how the performances of CT's models are affected by the choice of starting point.

2.1 *Ad hoc* decisions

It is now important to notice that there are already several, what we will call, *ad hoc* decisions made when considering the previous-best-forecast, the inverse pMSFE, and the shrinkage methods, while some of these decisions could perhaps better be motivated (or updated) by the data. We will now identify these *ad hoc* decisions and we will also sketch how these decisions could be updated by data by means of an automated algorithm. The details of the automated algorithm will be discussed in the Section 3.

First, when contrasting the previous-best method to the inverse pMSFE method, we notice that the former method only assigns a weight to one forecaster while the latter method gives a weight to all eligible forecasters. However, the optimal number of forecasters with non-zero weighs might be somewhere between one and all. Experts could be ranked based on their pMSFE scores, and the combination of best-ranked experts that results in the lowest pMSFE could be selected.

Second, as concerning the inverse pMSFE method, we notice that only the forecasters with a 'sufficiently long track record' receive an individual pMSFE weight. Of course, the length of the track record could influence the score. Indeed, the longer the track record, the fewer the participants who receive a pMSFE-based score. On the other hand, a track record that is too short might have too little information from which to estimate a forecaster's ability. In other words, could there be an optimal choice for the length of the track record? Incidentally, rather than taking the largest common sample of forecasters who have submitted at least ten contiguous observations in the past, we will take as a track record of a forecaster the set of consecutive forecasts submitted since

the last available realization. The advantage of CT's approach is that more forecasters are eligible even when the length of the track record is large. The disadvantage is that the amount of common observations is not fixed, and this complicates the comparison of different track record sizes. It might even be that no observation is shared by all eligible forecasters, in which case the Mean SPF model is used. Finally, the common observations might concern the distant past, which could render them less relevant for assessing the forecaster's current forecasting ability.

Third, and regarding the shrinkage factor, we noted already that CT either use a shrinkage factor of 0.25, or a shrinkage factor of 1. It might very well be that the optimal shrinkage factor lies somewhere in between and that it even changes over time. In our analysis below, we shall use a simplified version of (5) given by

$$w^i = (1 - \phi)\hat{\omega}^i + \phi\frac{1}{N_t}, \tag{7}$$

where $0 \leq \phi \leq 1$ and $\hat{w}$ is the inverse pMSFE weight. The reason for using this simplified version is that ξ_t can be larger than 1 when $N_t > v_t$. Also, we aim to control for the size of the sample and the number of forecasters by means of the automated algorithm. Given (7), the question now becomes how to optimally choose the shrinkage factor ϕ. When the shrinkage factor is optimized by minimizing the pseudo-out-of-sample prediction errors, a discordance between weights that are estimated using the estimation sample and the hold-out sample is effectively penalized.

Fourth, CT do not consider the forecasting errors that experts made for other variables when computing pMSFE-based scores. The main problem with finding reliable estimates for expert forecasting ability is that not enough data is available as a result of the discontinuous response rates, so adding relevant data might be worthwhile. An expert's ability to forecast NGDP might be informative about her ability to predict PGDP, for example. If the expert

submitted a forecast for PGDP, she will also have done so for NGDP, so adding her forecast error of NGDP does not alter the number of eligible candidates. Under the assumption that forecasting errors are of similar size on average, we can add these forecasting errors together to compute inverse MSFE score. The variables that are included for computing inverse MSFE scores are called 'score variables.' Based on pseudo-out-of sample forecasts, we can decide whether it is helpful to include NGDP forecast errors when pooling expert forecasts of PGDP.

Fifth, CT only take the pMSFE scores for an h-period-ahead forecast when based on the h-period-ahead individual forecast errors. As mentioned earlier, we might just as well use $\rho-$period-ahead individual forecast errors to determine the pMSFE-based weights when producing $h-$period-ahead forecasts. To avoid confusion, the horizons that are used to compute pMSFE scores will be referred to as 'score horizons'. The reason for allowing ρ to differ from h, is that, for example, $(h+1)$-period-ahead forecasts are generally more difficult to create than h-period-ahead forecasts. Hence, such forecast errors could be more useful in selecting the best forecasters for h-period-ahead forecasts. On the other hand, the latest available $(h+1)$-period-ahead forecast error will be less recent than the latest available (h)-period-ahead forecast error, so that the latter might perhaps be more informative about the forecaster's *current* forecasting ability. The length of the optimal track record and the number of included forecasters will similarly be influenced by the chosen score horizons. If the reader thinks it is odd to use different score variables or score horizons when pooling experts or models, then we would like to point out that scientists often select models based on their performances in various studies as documented by the scientific community. Note that the forecast errors of different horizons cannot simply be added together to determine MSFE based scores. On the one hand, because shorter horizons will have smaller forecasting errors on average than longer horizons, which decreases their influence on the MSFE score. On the other

hand, because some score horizons might be more relevant than others. Hence, one needs to give weights to different horizons when computing MSFE scores. One might use an alternative evaluation function so that scores across different score horizons are better comparable (MdRAFE below). The decision on which score horizons to include could be based on pseudo out-of-sample performance. The advantage of this approach is that one creates more data on which to base expert ability. The disadvantage is that the MSFE evaluation function cannot be used. Another solution is to combine expert forecasts based only on ($\rho = 1$)-period-ahead prediction errors, and combine expert forecasts once more using only ($\rho = 2$)-period-ahead prediction errors, and so on; and to select the best combination of score horizons afterwards based on their pseudo out of sample performance. This is just what we will do. Not only can an MSFE evaluation function still be used, but the total amount of perturbations in the algorithm will also be reduced considerably.

Sixth, CT look at MSFEs for evaluating forecasters and root mean squared forecast errors (RMSFE) to evaluate the forecasting accuracy of their models. If forecasting accuracy is defined in terms of RMSFE, then it might be interesting to look at forecasting combination weights based on RMSFEs next to MSFEs. Taking the root of the MSFEs will shrink the inverse weights towards the mean and the monotonic transformation will have no effect on the ranking of experts. Of course, many other evaluation functions could have been used, and MSFE is yet another example of an *ad hoc* choice. We therefore shall add an evaluation function based on the Relative Absolute Forecasting Error (RAFE) loss function by taking the median of

$$RAE^{i}_{t,h} = \frac{\max(\mid \hat{Y}^{i}_{t,h} - Y_t \mid, 0.1)}{\max(\mid \bar{Y}_t - Y_t \mid, 0.1)}, \tag{8}$$

see (Hyndman and Koehler, 2006). The median relative absolute forecasting error will get the acronym 'MdRAFE'. There are various differences between

squared forecasting error (SFE) and RAFE loss functions. In RAFE, absolute errors are used instead of squared errors in order to downplay the influence of large forecasting errors. Further, the absolute forecast errors are divided by the absolute forecast errors of the mean survey forecast. This makes forecast errors more comparable across time, across score horizons and across score variables. A minimum absolute forecast error of 0.1 is used to avoid extreme values. This restriction will also be applied when computing inverse pMSFE or pRMSFE weights, if only because prediction errors are sometimes zero. Depending on the analyst's goals, like having consistent and low forecasting errors, one accuracy measure might be preferred to another in evaluating the final forecast. Nevertheless, multiple evaluation functions could be used in the process of finding an optimal final forecast in terms of the desired evaluation function. Somewhat related, Hendry and Krolzig (2005) look at p-values, SIC, and other diagnostic tests to select variables. Now the question becomes whether we can think of a way to optimize over the accuracy measure that is used in selecting and weighting individual and model forecasts. In all the steps of the algorithm, configurations are selected based on some evaluation function. That is why the selection of evaluation functions will be the final step in the algorithm.

Seventh, the starting point of the treatment sample in the expanding window setup might influence empirical results. This issue could influence all of CT's proposed models to beat Mean SPF. We shall consider treatment samples which begin in 1968Q4 rather than in 1981Q3. Not only do many of the variables of interest have structural breaks in their level and volatility, also the forecasts of SPF members can be discontinuous. For example, for the quarterly change of the GDP price index (PGDP), the mean expert forecast was biased downwards between 1968Q4 and 1981Q1 and biased upwards between 1981Q2 and 2011Q2. Hence, we would expect that starting points around 1981Q1 shall be used once they are available, unless information about earlier crises becomes relevant. The number of expert forecasts in a given quarter ranges from nine to eighty-two

with an average of thirty-eight. Also, some changes in submission deadlines were introduced when the National Bureau of Economic Research (NBER) started conducting the survey in 1990Q2 instead of the American Statistical Association (ASA). So, starting points after 1981Q1 might also be expected. Moreover, there are some caveats on using the individual identification numbers (Stark, 2012). Due to all these developments in the variable of interest and the survey itself, we would suggest that it might be sensible to consider different starting points of the treatment sample.[4] As was explained above, PT (2007) either use the single best starting point or a weighted average of all of the starting points based on a window of evaluating starting points of either ten or twenty. The optimal number of top ranked starting points to be selected might therefore lie between only the single best starting point and all the starting points. To find out, starting points could be ranked based on their pMSFE scores, and a combination of best-ranked starting points that results in the lowest pMSFE could be selected. Furthermore, rather than selecting a window of ten or twenty, we will take an average of multiple windows. Lastly, it will be shown how to account for the fact that an optimal starting point of the treatment sample might be different for different items.[5]

Next to identifying and updating *ad hoc* statistical decisions, one might try to anticipate how choices of configurations (of the empirical analysis) interact. Predicting which configurations will lead to the lowest forecasting errors is complicated by the fact that many different kinds of distributions of forecasters could be envisaged. For example, if there is little systematic difference in the

[4]Another issue with the survey is that the SPF did not collect four-quarter-ahead forecasts about 1970Q1-Q3, 1971Q1, and 1975Q3 (all part of treatment sample). To deal with this problem, we will use the associated three-quarter-ahead forecasts for four-quarter-ahead forecasts as well, when the latter were not collected.

[5]To our taste, an equal weighted mean of all windows should not be compared to an equal-weighted forecast, as PT do (Pesaran and Timmermann, 2007, pp. 144). When an equal-weighted mean is taken of all the starting points, the influence of observations gets larger as the observations becomes more recent. Early observations are discarded when more recent starting points are used. In that sense, the first-available forecast is more like an equal-weighted forecast, since all observations will have the same influence on the estimation process.

ability of forecasters, we might expect many forecasters to be included, along with a small track record (in order to exclude as few people as possible). When forecasting qualities of some forecasters are consistently better than those of others, then the number of forecasters might become lower and the track record higher (to get proper individual scores). It could also be the case that all forecasters consistently perform reasonably well, except for a small group whose performance measured over a relatively long time is poor. In that case, the number of forecasters, the length of the track record, and the degree of shrinkage towards the mean are high.

Although straightforward conjectures about the optimal combination of configurations are hard to make, one might hypothesize about how much the decision of one configuration influences predictions conditional on the choices made about other settings. For example, the track record, the score horizon, the score variables, and the accuracy measure will have more effect on forecasting accuracy when the number of forecasters is low. If many forecasters are selected anyhow, then the correct ranking based on these four items becomes less important. The track record, the score horizon, the score variable, and the accuracy measure also affect the weights assigned to the forecasters. Hence, when the degree of shrinkage towards the equal-weighted mean is high, the influence of these features on forecasting accuracy is expected to decrease. Such information can be relevant in restricting the number of configurations that are considered by the algorithm.

2.2 Conclusion

In sum, when selecting and combining forecasters based on their previous performance, CT made *ad hoc* judgments about the length of the track record, the number of forecasters included, the degree of shrinkage towards the mean forecast, the score variables and score horizons used in computing individual scores, the starting point of the treatment sample, and the accuracy measures

used for combining and evaluating forecasts. We will address these decisions in the next section and discuss how they can be included in an automated algorithm to overthrow or robustify their conclusion that the mean survey is hard to beat.

3 Automated Algorithm

Our general automated algorithm called 'AA' primarily builds on the previous-best forecast, the inverse pMSFE method, and the shrinkage model. The algorithm that builds on SIC-bias will be called TCOMB-SIC-bias. There are two main procedures that are used for combining configurations. The first is called 'COMB', and it combines forecasts using shrunken inverse (RMSFE, MSFE, MdRAFE)-scores as weights. The second algorithm, called 'TCOMB', pools forecasts with different starting points of the treatment sample with the help of COMB. We will start by introducing these two sub-algorithms.

3.1 COMB-algorithm

Algorithm 1: COMB-Algorithm

input : Forecasts of models and indication of evaluation function and starting point.

output: Combined point forecasts with shrunken inverse score weights.

a. Rank input models based on their scores.

for *Number best ranked models = 1:1:total number of models* **do**

  for *shrinkage rate = 0:0.1:1* **do**

    b. Compute combined forecast based on shrunken inverse score weights for a given number of best ranked models and a given shrinkage rate.

  end

end

c. Select the **b**-model forecast with the lowest pseudo out-of-sample score.

A simple strategy to optimize over a single configuration with an automated algorithm is the following. First, the analyst defines a set of possible values of the configuration. Second, pseudo-out-of-sample forecasts are recursively created for each possible value. Third, pMSFEs are computed for each forecast, and the configuration value with the lowest pMSFE is selected.

Rather than selecting the single best configuration, we could also attempt to optimize over the number of best ranked configurations to be amalgamated. In determining how many best-ranked models should be averaged, for example, the analyst starts by defining the set $N = \{1, 2, \ldots, 20\}$ of the possible numbers of models that are joined. Pseudo-out-of-sample forecasts are subsequently produced by taking the single best-ranked model, the mean of the two best-ranked models, and so on. A forecast is finally made with the number of best-ranked models that led to the lowest pMSFE. Note that we only look at combinations of best-ranked models instead of all possible combinations between models in

order to reduce the total number of perturbations.

Furthermore, instead of taking the equal-weighted mean of a number of models, one might also use shrunken inverse pMSFE weights as in (7). In that case, the analyst would have to optimize over two sets of restrictions. One straightforward way to do so is for the analyst to define the set of possible shrinkage values, say $\phi = \{0, 0.1, 0.2, \ldots, 1\}$, and merge that set with the set concerning the included number of models. In this case, 220 different perturbations are computed by changing the shrinkage factor and the number of included models, and an out-of-sample forecast is produced using the settings with the lowest pMSFE. These steps will be called 'COMB' and are summarized in Algorithm 1. The notation (0:0.1:1) stands for a set going from 0 to 1 with increments of 0.1. Note that a stronger prior might be obtained by taking a weighted average of the outcomes in step **c** and the analyst's preferred number of models. The analyst might also take a weighted average between the scores assigned to each perturbation and the pMSFE scores.

3.2 TCOMB-algorithm

Algorithm 2: TCOMB-Algorithm

input : Forecasts of models, indication of evaluation function and backup model

output: A combined point forecast with shrunken inverse score weights.

for *Window = 10:1:30* **do**

  for *Starting point = 1968Q4:Q:2003Q3* **do**

    a. Use **COMB** to combine input models for a given starting point and evaluation function.

  end

  b.

  if *Sufficient pseudo out-of-sample observation in window* **then**

    Use **COMB** to combine starting points for a given evaluation function and window size.

  else

    Use backup model.

  end

end

c. Take equal-weighted mean of **b**-models

COMB, the algorithm that selects and combines models, can in turn be used to select other configurations. The analyst might want to optimize over the starting point of the treatment sample in an expanding window setup. A set of possible starting points can be defined and pseudo out-of-sample forecasts can be computed using different starting points. Only forecasts based on at least twenty observations are considered. COMB can then be applied to select the optimal combination of starting points by assigning shrunken inverse pMSFE weights to a number of best-ranked starting points. The next question becomes how large the moving window should be for evaluating the performance of the starting-point-dependent models. A window that is too small is unreliable, and a window that is too large excludes more recent starting points. One might vary

the amount of window sizes, say, from ten to thirty and simply take an equal-weighted mean of the resulting forecasts. If, at the start of the sample, sufficient pseudo-out-of-sample forecasts for the starting-point-dependent models are not yet available for the smallest window, the forecasts of a 'backup' can be used. Such a backup could be the forecasts generated by the earliest available starting point. We do not use the average of the starting points which are available as a backup, because we have no *a priori* reason to assume that earlier data points are less informative than later data points. Early observations are discarded when the starting point is increased. The steps enumerated in this paragraph, which are aimed at finding the optimal combination of models and starting-points, are called 'TCOMB' and are summarized in Algorithm 2.

To investigate how the performance of SIC-bias was affected by the sample chosen, we will apply TCOMB to SIC-bias as well. In the first step, SIC-bias is used instead of COMB to select Mean SPF or the bias-adjusted model for a given starting point and an MSFE evaluation function. This model will be called 'TCOMB-SIC-bias'. SIC-bias will also be performed using the PT setup, whereby the single best starting point was selected or a weighed average was taken over all available starting points using windows for evaluating the starting points of ten and twenty. The backup model, used when there are not sufficiently many pseudo-out-of-sample forecasts to compare starting points, relies on the earliest available starting point.

3.3 AA-algorithm

Algorithm 3: AA-Algorithm

input : Expert forecasts and realizations

output: AA forecasts

for *Evaluation function* $\in$ {*RMSFE, MdRAFE, MSFE*} **do**

  for *Starting point = 1968Q4:Q:2003Q3* **do**

    for *Score horizon = 0:1:4* **do**

      a. Compute forecasts with varying track records (1:1:10), number of best ranked experts (1:1:40), shrinkage rates (0:0.1:1), and score variables (y_1, y_1 & y_2); for a given score horizon, starting point, and evaluation function.

    end

    b. For a given evaluation function, select optimal **AA.a** configurations for each score horizon, add Mean SPF and use **COMB** to combine these six models.

  end

  Continue with **TCOMB.b** and **c** to combine starting points.

end

c.

for *Evaluation function* $\in$ {*RMSFE, MdRAFE, MSFE*} **do**

  Apply TCOMB to combine **AA.b** models for a given evaluation function.

end

Take the equal-weighted mean of the resulting three models.

Algorithm 3 presents the general automated algorithm 'AA' to be employed in our approach. The main procedure is as follows. In step AA.a pseudo-out-of-sample forecasts are created based on all kinds of combinations between different items for a given score horizon and evaluation function. In step AA.b, the best set of configurations is selected for each score horizon after which the optimal combination of score horizons is determined, for a given evaluation

function. Finally, in step AA.c, the optimal combination of evaluation functions is determined. The models at the end of steps AA.a and AA.b will be called AA.a models and AA.b models respectively.

In the first step, called AA.a, every possible combination between the configurations of four items are generated for each score horizon and evaluation function. These four items are the number of experts that are included, the length of the track record, the shrinkage level, and the score variables. The set of possible numbers of best-ranked experts is $N = \{1, 2, \ldots, 40\}$. If the number of forecasters is restricted to be twenty-five and only ten experts are eligible, then these ten forecasters will be used instead. The length of the track record varies from 1 to 10. The degree of shrinkage varies from 0 to 1 with increments of 0.1. Regarding the score variables, the prediction errors of PGDP (y_1) or PGDP and NGDP (y_1 & y_2) are used when pooling PGDP (y_1) forecasts. When both score variables are included, a pMSFE-score is computed by taking the mean of the squared forecast errors of both PGDP and NGDP, for a given track record, shrinkage rate, etc. The total amount of perturbations between these four items is $40 \cdot 10 \cdot 11 \cdot 2 = 8,800$ for each score horizon and evaluation function.

At the start of AA.b, the set of configurations with the best pseudo out-of-sample forecasting accuracy is selected for each score horizon based on a given evaluation function. There are five score horizons, namely $\rho = 0, 1, \ldots, 4$. The equal-weighted mean forecast is added to the five score-horizon-dependent models, in order to include forecasters that were not eligible due to their track record. To select the optimal AA.a configurations and combine the resulting six models, it will be determined which weight should be assigned to which starting point of the treatment sample. To this end, TCOMB is applied. Given a moving window of twenty observations to evaluate starting points and a pRMSFE evaluation function, we might find, for example, that it is optimal for an ($h = 2$)-period-ahead forecast in 1999Q3 to take only 1971Q1 as a starting point, whereby the

selected AA.a models of the $\rho = 2$ and $\rho = 3$ score horizons of step AA.a are combined with Mean SPF using full shrinkage.

Going into details about the backup model of AA.b, we do not use the first available starting point as a backup model in case there are no sufficient pseudo-out-of-sample observations to select a starting point, because the AA.a models produce their first forecasts at different points in time. For convenience we denote the first starting point, 1968Q4, as $q = 0$, and 1969Q1 as $q = 1$, and so on. In AA.b, starting points will be compared for a given window size when the $(q = 14 + h)$ starting point has a sufficient amount of known pseudo-out-of sample observations. When this starting point does not have enough observations for any of the windows, the forecast of the highest available starting point below $(15 + h)$ is used as a backup. The AA.a model with the largest amount of observations required to produce a forecast has a track record of ten, a score horizon of $\rho = 4$, and a forecasting horizon of $h = 4$. The first $(\rho = 4)$ forecast is known in real-time at $q = 5$; and the tenth $(\rho = 4)$ forecast is known at $q = 14$. At $q = 14$, the weights based on the previous ten observations are used to produce an $(h = 4)$-period-ahead forecast concerning $q = 18$. This forecast results in the first pseudo-out-of-sample forecast error for this model, which is available in the next quarter. Hence, at the eighteenth observation, all AA.a models will have produced an $(h = 4)$ period-ahead-forecast, conditional on the forecasters submitting enough contiguous predictions. If a set of restrictions in AA.a leads to one or more missing observations in the treatment sample, then it is excluded. The length of the track record and the score horizon influence the number of missing observations.

When we arrive at step AA.c, we have performed steps **a** and **b** for each of the three evaluation functions, namely RMSFE, MdRAFE, and MSFE. Each time models or expert forecasts were combined, one of the three evaluation functions was used. Which accuracy measure should now be used to compare the performance of the RMSFE, MdRAFE, and MSFE based forecasts? We shall

use each evaluation function to amalgamate the three models with TCOMB and take the equal-weighted mean afterwards. In case there are not enough pseudo-out-of-sample prediction errors to compare starting points in TCOMB, the mean of the three forecasts will be used as a backup.[6]

3.4 Conclusion

We have presented the algorithms to be used in this paper. The main procedures used for selecting configurations are COMB and TCOMB. COMB combines forecasts using shrunken inverse (RMSFE, MSFE, MdRAFE)-scores as weight. TCOMB pools forecasts with different starting points of the treatment sample with the help of COMB. The items of AA are updated in three steps. In step AA.a pseudo-out-of-sample forecasts are made based on all kinds of combinations between different items for a given score horizon and evaluation function. In step AA.b, the best set of configurations is selected for each score horizon after which the optimal combination of score horizons is determined, for a given evaluation function. Finally, in step AA.c, the optimal combination of evaluation functions is determined. TCOMB-SIC-bias arises when TCOMB is applied to SIC-bias. It should be stressed that at any given time in the algorithm, only those flash realizations are used in AA that were also available to the experts when they submitted their forecasts.

4 Tools for Analyzing the Automated Algorithm

In the previous section it was discussed how AA combines configurations with the number of experts, the length of the track record, the shrinkage rate, and the score variables in AA.a, how it pools the score horizons with the Mean SPF model in AA.b, and how it selects the accuracy measures in AA.c. The reader might start to worry about the number of items that are included. With so

[6] An alternative approach would be to use the same evaluation function for comparing the three forecasts as the one that is used for evaluating the final forecasts.

many candidate configurations the algorithm could quickly turn into a black box. How do we know which configurations were selected, how successful the algorithm was in selecting the right configurations, and which configurations resulted in good or bad forecasting accuracy? On the basis of such information, the analyst might want to consider more configurations of particular items (a new evaluation function for instance), or she might want to included an informative prior about configurations to avoid risky configurations getting high weights. Where the set of configurations of some items might be enlarged, other sets of configurations can be abridged. It is important to find out which configurations are irrelevant in order to increase the speed of the algorithm.

4.1 Evaluating AA

The success of AA in selecting the right set of configurations depends on whether the analyst defined the right set of configurations, whether there is a set of configurations that is consistently better than others for a reasonable period of time and on whether AA employs the right strategy to find this set. Tools will be presented below to study these aspects. Here, we discuss how it can be established whether AA was successful in selecting the right set of configurations and how much uncertainty was involved in choosing a set of configurations. Importantly, this helps to robustify claims about the difficulty of beating mean forecasts.

To evaluate forecasts we shall follow Capistrán and Timmermann in using RMSFEs relative to the RMSFE of the Mean SPF model. That is, the RMSFE and MdRAFE scores are divided by the RMSFE and MdRAFE scores of Mean SPF. So, when this relative score is below 1, the model outperformed Mean SPF, and when the score is higher than 1, the forecasting accuracy of Mean SPF was better. When relevant, forecasting accuracy in terms of the MdRAFE and MSFE evaluation functions will also be presented. To save space, we will highlight the results on $\text{PGDP}_{h=2}$ when analyzing the dynamics of the algo-

rithm, for no other reason than that two is halfway zero and four. In case another horizon or variable is interesting, it will be presented as well.

AA's general performance can be assessed by comparing AA's RMSFEs (or other accuracy measures) to benchmark models, such as SIC bias and Mean SPF. AA's ability to select the right accuracy measure in step AA.c can simply be determined by comparing AA's RMSFE results to those of the three individual AA.b models and also to the RMSFE of the mean of the three AA.b models.

To study how successful the algorithm was in selecting AA.a configurations, the difference between the absolute prediction errors of AA and those of each perturbation in AA.a will be computed for each point in time. That is, we consider all the forecasts of the accuracy measures, score horizons, score variables, number of experts, track records, and shrinkage rates. The relative absolute errors of the 5^{th} and 95^{th} percent best set of restrictions will subsequently be selected to generate a ninety-percent interval plot. So, if the entire interval is above zero in the plot, this means that the absolute error of AA in that period is in the top five percent of all possible perturbations. When the middle of the interval at a given time is zero in the plot, AA performs similar to taking the median of all the forecasts of AA.a. AA has worse predictive accuracy than 95% of the perturbations if the entire interval is below zero. The 68% interval will also be plotted. A plot of such an interval also shows the uncertainty involved in selecting configurations. Note that the total amount of forecasting uncertainty is larger than this interval. Summary statistics will be presented to compare the performance of AA relative to taking the mean of all the perturbations in AA.a. A histogram will also be presented which shows the relative frequency with which AA was ranked in the top $10\%, 20\%, \ldots, 100\%$ best AA.a models.

4.2 Updated configurations

To turn an automated algorithm into a white box, we should start by studying which configurations were selected over time. This tells us which settings had good pseudo out-of-sample forecasts in the past, whether there is some consistency in the settings chosen, how many configurations of the same item are combined (for example, parsimony of selection of starting points), and how the choice of configurations is affected by structural changes in the underlying process (for example the economy, SPF methods, and so on).

Distribution images will be used to show how decisions are updated over time. Figure 1, which shows when experts submitted forecasts, is an example of such an image. To study how for example the choice of a starting point is updated over time, one starts by collecting the weights ($w^s_{m,h,t}$) that were given to the starting points (s) for each moving window size (m), for a given h-period-ahead-forecast at time t. Remember that various sizes of a moving window (10:1:30) are employed to evaluate the performance of the starting points and that each starting point receives a (shrunken inverse score) weight when it is selected. The accumulated weight $q^s_{h,t}$ assigned to a given starting point is given by

$$q^s_{h,t} = \sum_{m=10}^{M_t} w^s_{m,h,t} * 1/M_t, \tag{9}$$

where the maximum moving window size (M_t) increases up to thirty when more observations become available. The higher the weight of a given starting point, the darker is the mark in the image. The distribution images of the selection of other configurations are constructed in a similar fashion.

4.3 AA configurations and forecasting accuracy

With the help of a distribution image we can see which configurations have led to good pseudo out-of-sample forecasts over time. It remains unclear what the effect of a choice in configurations is on forecasting accuracy. How much larger

would prediction errors have been if ($\rho = 3$) instead of ($\rho = 2$) period-ahead predictions errors were used when pooling expert forecasts? As the algorithm contains a lot of perturbations, it is infeasible to tabulate the RMSFE scores for each model. This is why an accuracy image will be presented. On the basis of such an image, the analyst can ascertain how *ad hoc* decisions affect forecasting accuracy and which items could benefit from receiving more candidate configurations. The analyst can construct an informative prior which avoids poor performing sets of configurations being selected when little data is available.

The strategy we employ is to define a benchmark model, and to see how forecasting errors are affected by changing configurations in that benchmark model. The benchmark model R_0 is specified as having twenty forecasters, a track record of five observations, a shrinkage rate of 0.50, a score horizon of two, only PGDP as a score variable, and an RMSFE accuracy measure. Now, to find out how forecasting accuracy is affected by the choice of shrinkage rate, we might redefine the degree of shrinkage in R_0 to be 0, 0.1, ..., 1, and compute prediction errors for each alterations. An 'AE plot' can be made which shows the absolute errors across time of all configurations of a particular item. Such a plot might indicate whether some configurations work best, and whether predictions are affected by these configurations at all. An AE plot might become fuzzy when too many configurations are included.

To summarize the relation between configuration settings and forecasting accuracy, an accuracy image is constructed. We redefine the configurations of two items in R_0 and compute the RMSFE for these different settings. An accuracy image shows the RMSFE values when one item of R_0 is changed on the vertical axis, and another item of R_0 is changed on the horizontal axis. Only those observations are used for which all models in the plot produced a forecast. This way the forecasting accuracy of far more combinations of configurations can be analyzed. One might think of R_0 as an *ad hoc* model, and an accuracy image as a means to study the effect of *ad hoc* choices on forecasting accuracy.

4.4 Restricting items

The number of configurations per item can easily be restricted if we know how much predictions vary as a result of changing the configurations of that item. If all configurations of a particular item amount to the same prediction, it is of no use to compute all these perturbations. An AE plot gives an indication of how much predictions vary across different configurations, but such a plot quickly becomes too vague. When the mean variation across configurations of a particular item are summarized in an image, the importance of items can conveniently be studied for different reference models.

To measure the average variation in forecasts caused by changing a configuration of an item, we define

$$S_i(R_j^v) = \text{mean}_t\big(\text{std dev}_{w|t}(\hat{Y}_{t,R_i^w})\big), \tag{10}$$

where i refers to an item (number of experts, track record, shrinkage rate, and so on) and $w = 1, 2, \ldots, W$ refers to the w-th configuration value of a particular item in the reference model. Hence, for each point in time, the standard deviation in prediction errors across different configurations is calculated. The mean is subsequently computed for these standard variations over time to find the average amount of variation across different perturbations. When S_i is low, this means that forecasts are little influenced by an *ad hoc* choice made for that particular item. S will also be computed to find out what the average variation in predications are across different starting points and evaluation windows. Only those observations are used in computing S_i that are shared by all of the AA.a forecasts.

Various concerns might be expressed about a tentative interpretation of S_i as a measure of importance of a particular configuration choice. For one, S_i might not accurately capture average variation in predictions across configurations. The analyst can substitute squared errors by absolute errors, mean by median,

and so on. To mention three other impracticalities: S_i seems to depend heavily on which model is chosen as a reference model, S_i might average away large influences of certain settings, and S_i does not inform us about the interaction between different choices of configurations. These three related issues can be addressed by studying how S_i is affected by the choice of the initial reference model R_0. This is why the input variable in equation (10) is R_v^j. One might compute average variations in predictions across configurations of an item by varying the number of top-ranked experts in case the shrinkage rate of the reference model is zero ($S_{experts}(R_{shrinkage}^{\phi=0.0})$), and compute $S_{experts}(R_{shrinkage}^{\phi=0.1})$ once more when the shrinkage rate is 0.10, and so on.

An $S-$image will be presented whereby one item (j) of the reference model is changed on the vertical axis, and S_i is represented on the horizontal axis for all the items (i) in AA.a. With the help of this plot, an analyst can decide which items or combinations between items may receive less configurations. She might find, for instance, that the choice on the accuracy measure is only important when the shrinkage rate is low, since individual weights are based on accuracy measures. Less accuracy measures might then be considered for high shrinkage rates.

4.5 Conclusion

To summarize, by plotting the ranking of AA among all perturbations the performance of AA will be evaluated. Next, a distribution image will be used to show which weights were assigned to which configurations, and an accuracy image will be displayed that shows how forecasting accuracy is affected by adjusting configurations. Finally, the measure S will be tabulated or represented by an S-image to find out which items may receive less configurations.

5 Results

Following the same structure as the previous section, it will now be discussed how AA performed relative to benchmark models and the candidate models in AA, which configurations were selected by AA, how AA can be improved, and how AA can be restricted. Finally, it will be analyzed how the decision on the start of the treatment sample affects the performance of SIC-bias, by applying TCOMB to SIC-bias. At the end of this section, we will discuss whether AA and TCOMB-SIC bias improved the models on which they were based in terms of forecasting accuracy, and whether the claim remains standing that Mean SPF is difficult to beat.

5.1 Evaluating AA

Table 1 shows the relative RMSFEs and MdRAFEs of AA, TCOMB-SIC-bias and some benchmark models. AA will first be compared to Mean SPF, and subsequently it will be studied how successful AA was in pooling evaluation functions in step AA.c and in selecting the right configurations of all the AA.a models.

AA has worse forecasting accuracies in terms of RMSFE (and MSFE) than the Mean SPF for nearly all horizons of PGDP and NGDP. Only for $\text{PGDP}_{h=1}$ the relative RMSFE is below 1, namely .98. When looking at MdRAFE, AA performed slightly better for PGDP and slightly worse for NGDP. The MSFE outcomes have the same patterns as the RMSFEs, although the relative differences are larger. For $\text{PGDP}_{h=2}$, the relative MSFE of AA is .96, for instance, and for $\text{PGDP}_{h=3}$, the relative MSFE of AA is 1.11.

As can be seen from Table 1, the choice of the evaluation function can lead to quite substantial changes in forecasting accuracy. The relative RMSFE of $\text{PGDP}_{h=2}$ forecasts when configurations were pooled using an RMSFE evaluation is 1.12, 1.00 for MdRAFE, and 1.09 for MSFE. When we compare the

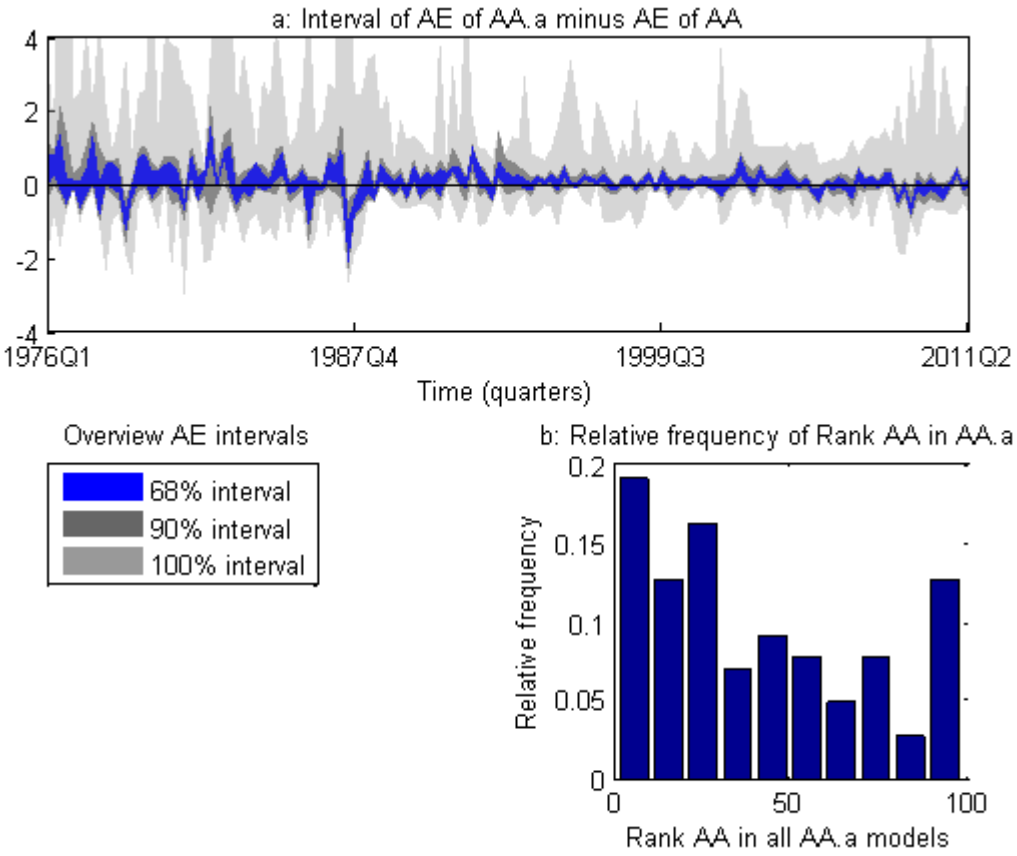
Table 1: AA results relative to Mean SPF (1968Q4:2011Q3)

Variable	Evaluation	Model	$h=0$	$h=1$	$h=2$	$h=3$	$h=4$
PGDP		AA	1.02	.98	1.03	1.05	1.03
		mean(AA.a)	1.03	1.02	1.06	1.03	1.04
		mean(AA.b)	1.00	.99	1.02	1.05	1.04
		AA_{rmsfe}	1.05	.99	1.04	1.12	1.09
		$\text{AA}_{\text{mdrafe}}$	1.02	1.02	1.07	1.00	1.03
		AA_{msfe}	1.01	.99	1.01	1.09	1.06
	MdRAFE	AA	1.00	.95	.97	1.00	.99
		mean(AA.a)	1.00	.99	1.02	1.00	1.01
		mean(AA.b)	1.00	.96	.97	1.01	.99
		AA_{rmsfe}	1.00	.99	1.00	1.02	1.00
		$\text{AA}_{\text{mdrafe}}$	.99	.94	1.00	.99	1.00
		AA_{msfe}	1.00	.98	.98	1.00	1.00
NGDP	RMSFE	AA	1.02	1.04	1.03	1.04	1.03
		mean(AA.a)	1.03	1.05	1.03	1.01	1.01
		mean(AA.b)	1.02	1.03	1.03	1.04	1.03
		AA_{rmsfe}	1.02	1.05	1.04	1.05	1.05
		$\text{AA}_{\text{mdrafe}}$	1.09	1.03	1.02	1.05	1.02
		AA_{msfe}	1.01	1.05	1.05	1.05	1.04
	MdRAFE	AA	1.02	1.03	1.01	1.00	1.00
		mean(AA.a)	1.02	1.02	1.00	1.02	1.02
		mean(AA.b)	1.03	1.04	1.01	.99	1.01
		AA_{rmsfe}	1.00	1.01	1.01	.99	1.03
		$\text{AA}_{\text{mdrafe}}$	1.01	1.02	1.01	1.00	1.00
		AA_{msfe}	1.00	1.03	1.01	.98	1.01

The RMSFE of the model is divided by the RMSFE of Mean SPF to obtain the results. The same goes for the other evaluation functions. So, when the result is smaller than one, the model outperformed Mean SPF; and vice versa. PT stands for the Pesaran and Timmermann setup for optimizing over the starting point of the treatment sample. The model mean(AA_a) takes the mean of all of the AA.a models. AA_{msfe}. The model mean(AA_b) takes the mean of the three evaluation functions instead of optimizing over the evaluation functions at step AA.c to obtain AA. The model AA_{msfe} shows the results when only $MSFE$ is used as an evaluation function. Similar for $\text{AA}_{\text{mdrafe}}$ and AA_{rmsfe}.

results of the three accuracy measures to AA, we can see that it is worthwhile to pool accuracy measures. AA is often ranked second best among the three accuracy measures. AA can also be compared to 'mean(AA.b)', which simply takes an equal weighted mean of the three accuracy measures. The performances of AA and mean(AA.b) are similar.

Figure 2: Interval of AE of AA.a minus AE of AA for PGDP



a. To find the 90% interval at a given time, the 5^{th} and 95^{th}% best set of restrictions are selected. The plot shows the absolute errors of forecasts of this interval minus the absolute errors of AA. So, if the entire interval is above zero in the plot, AA is in the top five percent of all possible perturbations in that period. The 68% and 100% intervals are constructed similarly.
b. This panel hows the relative frequency that AA was ranked in the top 10 %, 20 % etc. among all AA.a forecasts.

To study AA's ability to select the best AA.a configurations of all score horizons and evaluation functions, Figure 2 is presented. Figure 2.a shows the interval discussed in the previous section (4.1) of the absolute prediction

errors of AA relative to those of the models in AA.a for $\text{PGDP}_{h=2}$. In the period between 1976Q1 and 1987Q4 the forecasts of all the perturbations are particularly diverging. AA performed worse than 95% of the perturbations in 1979Q2, 1981Q3, 1987Q4, 1988Q1, 1998Q1, 1999Q3, 2005Q4, 2008Q4, and 2009Q2. AA was in the top 5% twelve times, the first time in 1980Q4 and the last time in 2011Q1. AA was nineteen times worse than 84% of the AA.a forecasts, and thirty-eight times better than 16%. Figure 2.b is a histogram of the relative frequency with which AA was ranked in the top 10%, 20%, ..., 100%, of the AA models. Even though AA's performance is not bad, it is clear that AA often did not manage to select the best set of AA.a restrictions. For other horizons of PGDP the ranking of AA is quite similar. The ranking of AA is more uniform for the horizons of NGDP, meaning that AA's rank was just as often in the top $10\%, 20\%, \ldots, 50\%$ as in the worst $10\%, 20\%, \ldots, 50\%$. Table 1 shows that the relative RMSFEs of AA compared to taking the mean of all the forecasts of AA.a. For both variables, AA is generally close to this benchmark.

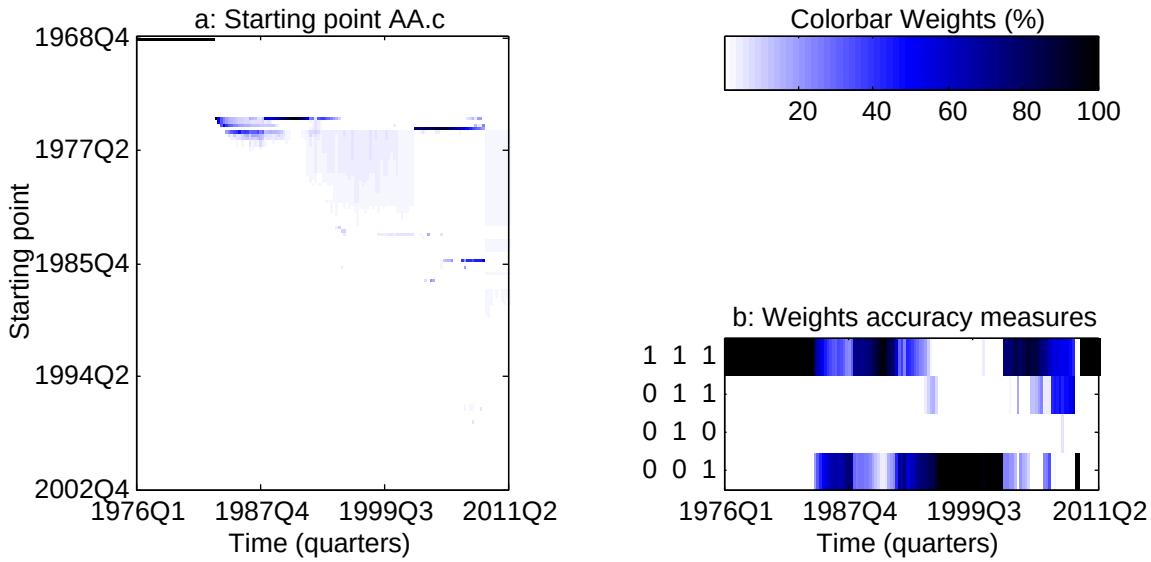
Although it is clear that AA failed to beat the Mean Survey, it remains unclear whether AA improved the models on which it was based and whether the equal-weighted mean can be beaten when configurations are defined and combined in a different way. To make some headway with the latter issue, we first need to turn the algorithm into a white box, by establishing how configurations were updated over time. It will subsequently be shown which configurations led to good or to bad forecasting accuracy in order to study the effect *ad hoc* choices have on forecasting accuracy, to see which items should receive more candidate configurations, and to be able to formulate an informative prior for future research. How is the forecasting accuracy of SIC-bias affected by the choice of the starting point, for instance, and which configurations could be added to further improve TCOMB-SIC-bias? Finally, we will identify which items could do with less configurations, so that the algorithms remains feasible even when promising configurations are added. These three matters will be

discussed for AA first, and for TCOMB-SIC-bias afterwards.

5.2 Updated AA configurations

To turn the automated algorithm AA into a white box, we start by presenting how the configurations of the items of AA were updated over time with the help of distribution images introduced in Section 4.2. The items are discussed in reverse order, because the selection of starting points and accuracy measures in AA.c also determines the selection of the starting points, score horizons, and other items in AA.b and AA.a.

Figure 3: Updated configurations AA.c for $\text{PGDP}_{h=2}^{\text{RMSFE}}$



a. This panel shows the weights that were given to the starting points of AA.c and AA.b respectively. As the colorbar indicates, the darker a dot, the higher the relative weight assigned to that starting point.
b. This panel shows the weight that was assigned to an evaluation function across time. Relative weights were accumulated over windows and selected starting points. The same colorbar applies.

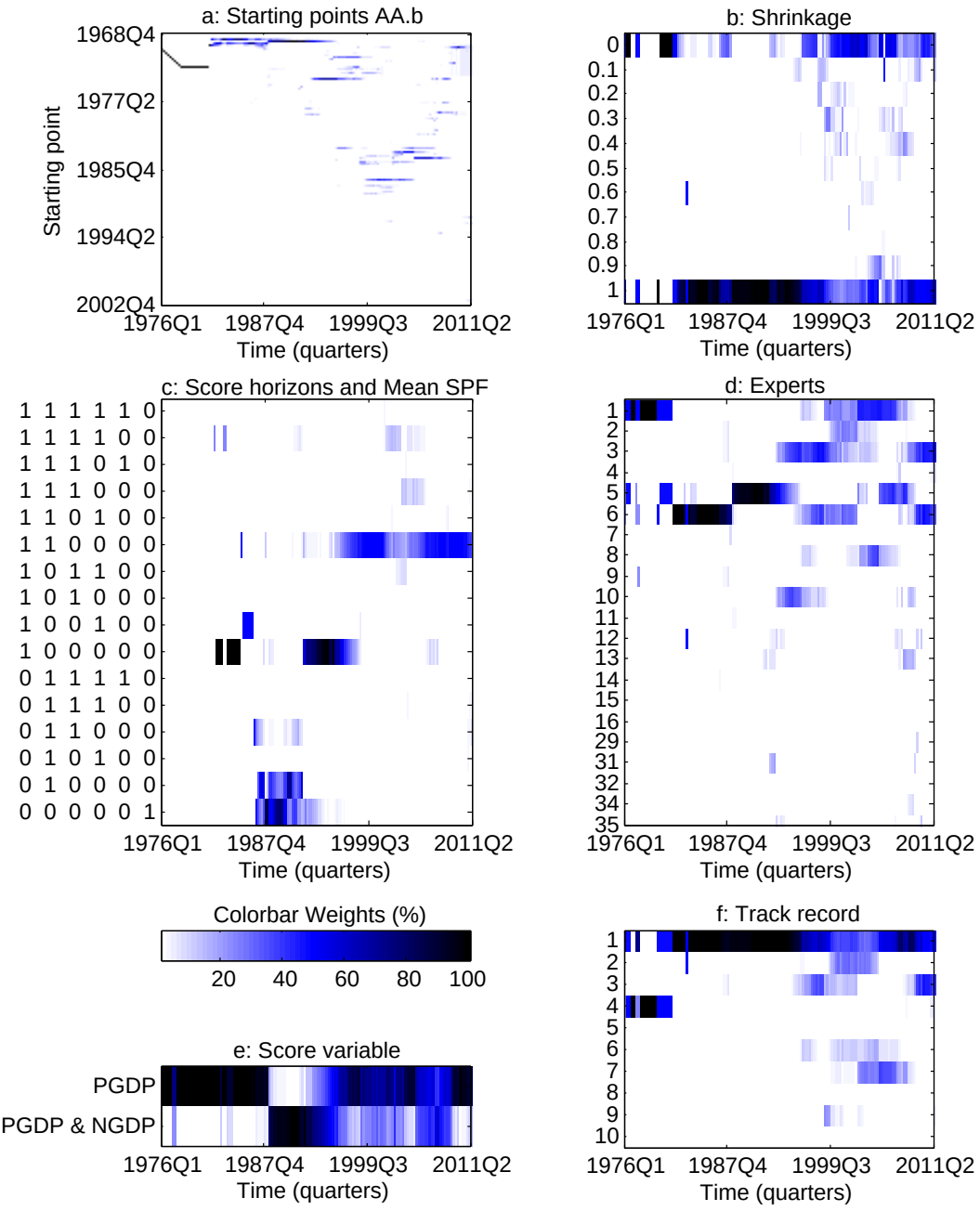
In step AA.c, the forecasts based on the RMSFE, MdRAFE, and MSFE evaluation functions were combined using TCOMB. Figure 3.a is a distribution image that shows which weights were given to which starting points in AA.c; for $\text{PGDP}_{h=2}$ and an RMSFE accuracy measure. Remember that the weight

is based on the weights each window gave to a particular starting point at a particular time. The darker a rectangle, the higher the accumulated weight was for that starting point. As of 1983Q3, the earliest starting point to be selected was 1974Q4. At 1995Q3, many starting points between 1974Q4 and 1982Q2 receive a positive weight. The starting point 1975Q3 is popular between 2002Q3 and 2007Q1, at which point 1985Q3 also comes into play. As of 2009Q2, starting points between 1975Q3 and 1984Q4 are nearly all given the same weight, because they all chose to combine the three evaluation function.

Figure 3.b shows the weights that were assigned to the three accuracy measures for $\text{PGDP}_{h=2}$ when using an RMSFE evaluation function for combining the accuracy measures. Weights are accumulated across the selected starting points of all windows. Generally speaking, either all three measures are pooled by a starting point, or only the MSFE accuracy measure is selected. Particularly between 1996Q1 and 2002Q2 the relative weight of the MSFE accuracy measure is high. To summarize the selection of evaluation functions for other horizons of PGDP; for $\text{PGDP}_{h=0}$, the RMSFE measure was often left out; for $\text{PGDP}_{h=1}$, each measures is selected at different time segments; for $\text{PGDP}_{h=3}$, MdRAFE was selected alone most of the time; and for $\text{PGDP}_{h=4}$, MdRAFE and MSFE were popular. Such variation between the weights assigned to evaluation function across horizons and time are also observed for NGDP.

In step AA.b, the best AA.a configurations are selected for each score horizon, after which the score horizons are combined with Mean SPF using COMB. This is done for all starting points. TCOMB.b and TCOMB.c are subsequently applied to combine the starting points. Figure 4 displays how the configurations of all these items were updated over time for $\text{PGDP}_{h=2}^{RMSFE}$. As Figure 4.a shows, the backup model was used until 1981Q2. Starting points increase up to 1987Q4 between 1993Q3 and 2004Q3, and return to the earliest starting points at the end of the sample to capture the turbulent dynamics of the seventies. When comparing Figure 4.a to Figure 3.a it is remarkable to find out

Figure 4: Updated configurations AA.a and AA.b for $\text{PGDP}_{h=2}^{\text{RMSFE}}$



These distribution images show the weights that were given to the configurations of an item over time for $\text{PGDP}_{h=2}^{\text{RMSFE}}$. As the colorbar indicates, the darker a dot, the higher the relative weight assigned to that starting point. Distribution images are shown for starting points (a), shrinkage rates (b), score horizons and mean SPF (c), experts (d), score variables (e), and track records (f). Regarding plot c, the score horizons $\rho = 0, \ldots, 4$ are the first five numbers, the last is Mean SPF; so (0 1 0 0 0 1) means that $\rho = 1$ and Mean SPF were selected.

that different items benefit from having different starting points of the treatment sample. For other horizons and accuracy measures, more moving window behavior is often observed, such that recent available starting points are also selected.

Turning to Figure 4.c, it is striking to observe that the ($\rho = 2$) forecast errors were rarely used for assigning weights to the expert's ($h = 2$) period ahead forecasts. Instead, $\rho = 0$ and $\rho = 1$ were generally employed. As the bottom row indicates, the Mean SPF had a better pseudo-out-of-sample forecasting performance than any AA.a model around 1987Q4. To summarize the selection of score horizons for the other forecasting horizons of PGDP, when $h = 0$, $\rho = 0$ and $\rho = 4$ are often selected. The Mean SPF forecast was again chosen around 1987Q4. For $h = 1$, most often the lowest three score horizons were merged and the Mean SPF was rarely included. When $h = 3$, the lowest three score horizons were again frequently incorporated, sometimes in conjunction with Mean SPF. For $h = 4$, first three score horizons were often used, and Mean SPF was selected around 1987Q4. When an MdRAFE measure is used, $\rho = 0$ is most popular for all forecasting horizons except $h = 3$, in which case $\rho = 1$ dominates. The Mean SPF is selected as often as in RMSFE.

Figure 4 b, d, e, and f are distribution images of the accumulated weights assigned to the different shrinkage rates, numbers of experts, score variables, and track record lengths respectively. That is, for each quarter and for each rolling window size, the weights were collected of the selected starting points and score horizons of AA.b, along with the weights given to the configuration of AA.a. Most of the times, between one and six experts are selected based on a track record of one, a PGDP score variable, and a high degree of shrinkage. Where selections of the length of the track record, the shrinkage rate, the score variables, and the score horizons are quite stable over time, the choice on the number of top-ranked experts to include varies quite a lot. To mention general characteristics of the AA.a specifications of other forecasting horizons; we ob-

served that the number of experts generally varies between one and eighteen and that the track record is often between one and three, although high track records are popular at the start and at the end of the sample. Both PGDP and NGDP are often used as score variables, except when the score horizon is $\rho = 1$, in which case PGDP is mostly selected. Generally speaking, full shrinkage is used when pooling expert forecasts; except for the score horizons $\rho = 4$, which also uses zero shrinkage quite often. When MdRAFE is used as an accuracy measure, far more experts are generally combined with a high degree of shrinkage, based on eight observation or less, and oftentimes only PGDP as a score variable.

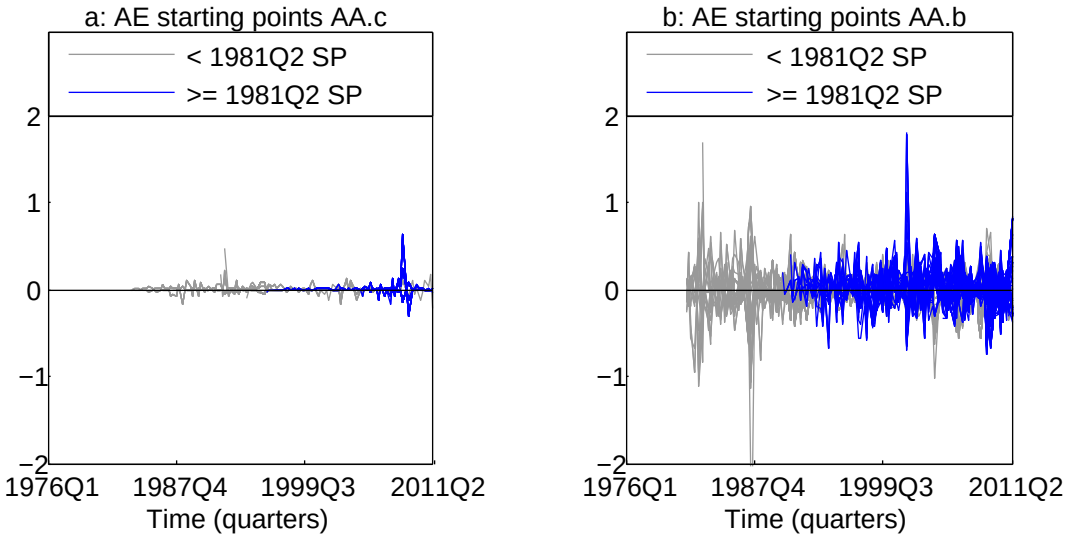
To conclude this subsection, we note that with the help of distribution images we have shown which configurations of AA were selected over time. TCOMB generally led to a parsimonious combination of starting points which was responsive to changing economic circumstances. Based on their pseudo-out-of-sample performance, it appeared to be beneficial to use low forecasting horizons, less than thirteen experts, a short track record, a high degree of shrinkage, and sometimes also NGDP as a score variable. Although these configuration settings changed across forecasting horizon, accuracy measure, and dependent variable, they did not display overly erratic behavior.

5.3 AA configurations and forecasting accuracy

To improve AA, we need to know which sets of configurations lead to bad forecasting accuracy and which sets of configurations lead to good forecasting accuracy. On the basis of this information, researchers can be motivated to investigate new configurations of a particular item and they can construct an informative prior. Moreover, the effect of the *ad hoc* choices made in CT's previous-best forecast, inverse pMSFE method, and the shrinkage model on forecasting accuracy can be examined.

Commencing with the absolute errors of different starting points of AA.c

Figure 5: Absolute errors of starting points AA.c and AA.b for $\text{PGDP}_{h=2}^{\text{RMSFE}}$



a. This panel shows the weights that were give to the starting points of AA.c and AA.b respectively. As the colorbar indicates, the darker a dot, the higher the relative weight assigned to that starting point.
b. This panel shows the weight that was assigned to an evaluation function across time. The same colorbar applies. Relative weights were accumulated over windows and selected starting points. The three numbers on each row of the vertical axis correspond to RMSFE, MdRAFE, and MSFE respectively. For instance, (0 1 1) means that RMSFE was not selected while MdRAFE and MSFE were selected.

and AA.b, Figure 5.a shows the absolute error of a particular starting point at a particular time minus the absolute error of taking the mean of all the starting points at that time for an evaluation window of size twenty. When this difference is negative, that particular starting point outperformed the average forecast of all the starting points. At the end of the sample, starting points past 1981Q2 are more volatile. Figure 5.b indicates that absolute prediction errors vary quite a lot as a result of choosing a different starting point when combining AA.b models. There appears to be no clear difference between starting before and after 1981Q2.

If someone were to choose the best configurations after observing the 'out of sample' performance of all AA.a perturbations, what *ad hoc* choices would he or she have made? Well, there is no set of restrictions which outperformed Mean SPF for all horizons of PGDP or for all horizons of NGDP. Table 2 shows what the best and the worst set of restrictions are for each horizon of PGDP and NGDP. For some items, the optimal configurations vary a lot, such as the accuracy measure, score variable, and shrinkage rate. The number of experts ranges from four to twenty-three for PGDP, the track record is short, and score horizons $\rho = 4$ and $\rho = 5$ are not selected. The best PGDP models for MdRAFE contain between two and twelve experts, a track record lower than six, and a score horizons of $\rho = 2$ or lower. As Table 2 shows, the best model for NGDP based on relative RMSFEs contains between three and six experts and a track record of five or less. Score horizon $\rho = 3$ is not within the best models. Although the scores of these models did not arise by selecting the best set of AA.a configurations at each point in time, but by selecting the overall best set of AA.a configurations, it is striking to observe that the relative RMSFE are so close to one. When AA.a configurations are optimized *a posteriori* for each point in time, one notices that it is often optimal to select the single best expert only. The problem is that the single best expert models are also ranked worst frequently. Indeed, Table 2 shows that to get the poorest overall results, one

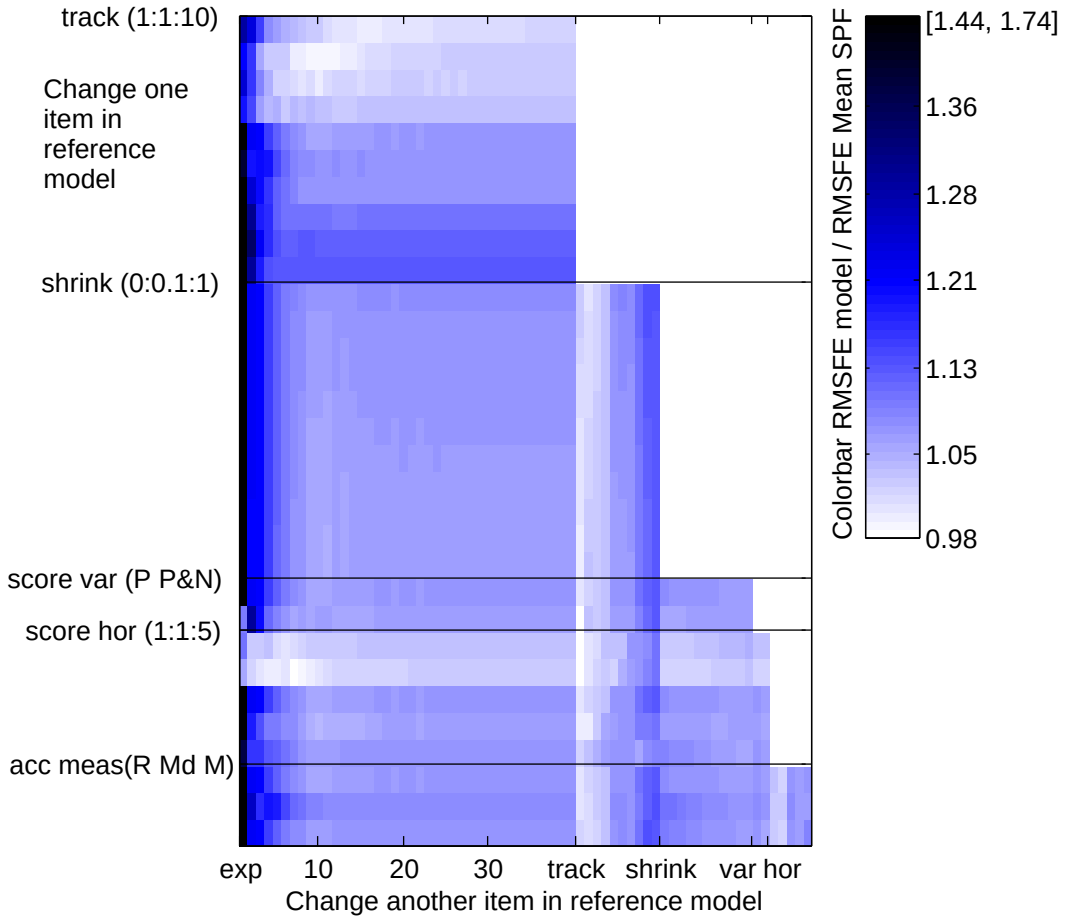
Table 2: Best and worst *ad hoc* AA.a models (1968Q4:2011Q3)

variable		h=0	h=1	h=2	h=3	h=4
PGDP	Best RMSFE	0.97	0.91	0.97	0.95	0.95
	Experts	23	4	13	13	18
	Track record	1	3	1	1	1
	Shrinkage	0.5	0.9	1.0	0.8	0.0
	Score variable	2	1	1	1	2
	Score horizon	3	1	2	2	3
	Accuracy measure	2	2	1	1	3
	Worst RMSFE	2.09	1.60	1.88	1.89	2.05
	Experts	1	1	1	1	1
	Track record	1	1	2	2	4
	Shrinkage	-	-	-	-	-
	Score variable	2	2	1	1	2
	Score horizon	5	2	5	5	1
	Accuracy measure	1	1	2	2	1
NGDP	Best RMSFE	0.94	0.99	0.98	0.93	0.95
	Experts	6	6	6	3	5
	Track record	1	1	1	3	5
	Shrinkage	1.0	1.0	1.0	0.0	0.0
	Score variable	2	2	2	1	1
	Score horizon	4	5	1	2	2
	Accuracy measure	1	2	1	2	3
	Worst RMSFE	2.01	1.33	1.30	1.43	1.67
	Experts	1	1	1	1	1
	Track record	1	4	2	2	2
	Shrinkage	-	-	-	-	-
	Score variable	1	2	2	2	2
	Score horizon	5	1	5	4	3
	Accuracy measure	2	1	2	2	1

This model shows the *a posteriori* best and worst *ad hoc* choices. The RMSFE of the model is divided by the RMSFE of Mean SPF to obtain the results. So, when the result is smaller than 1, the model outperformed Mean SPF; and vice versa.

needs to select only the single best expert.

Figure 6: Influence configurations on forecasting accuracy for $\text{PGDP}_{h=2}$



Depicts relative RMSFEs when two items in the reference model R_0 are varied. The RMSFE of the resulting model is divided by the RMSFE of Mean SPF, so a value lower than one means that the model outperformed Mean SPF. As the colorbar indicates, a darker rectangle corresponds to worse forecasting accuracy.

Although Table 2 gives us some indication of which set of configurations leads to good or to bad performances, it does not inform us what the influence of specific items are on forecasting accuracy. To this aim, Figure 6 is presented. Remember that an initial reference model R_0 was defined before in section 4.3, which was composed of ten experts, a track record of five observations, a

shrinkage rate of a half, PGDP as a score variable, a score horizon of two, and an RMSFE accuracy measure; see Table 4. One item of R_0 is altered on the vertical axis of the accuracy measure, and a configuration value of another item is altered on the horizontal axis. For twelve observations there was some model that did not produce a forecast, so these observations were left out. When the score horizon in the initial reference model R_0 is changed into $\rho = 0$ and the length of the track record is one instead of five, a relative RMSFE of 0.98 is found. The rectangle associated with these adjustments to R_0 is completely white, because it is the lowest score in the accuracy image. The worst RMSFE occurs when the track record is changed into nine observations and the number of top ranked experts is changed into one. This rectangle is black, and represents the number 1.74. In fact, when the other items are varied in R_0, the use of the single best forecaster often leads to poor forecasting accuracy. Higher forecasting accuracy is achieved when between four and fourteen experts are included. There appears to be little difference in forecasting accuracy when twenty top ranked experts are pooled instead of forty. It is also clear from Figure 6, that lower track records result in higher forecasting accuracy. When few experts are included and the shrinkage rate is low, a track record of one has worse forecasting accuracy than a track record of two, other things equal. Higher shrinkage rates increase forecasting accuracy with the slightest degree when the reference model is altered across number of experts, track records, and accuracy measures. The relative RMSFEs are often lower when both PGDP and NGDP are used as score variables. The score horizon has a strong effect on forecasting accuracy. Lower score horizons clearly outperform higher score horizons. Even when only one expert is used, $\rho = 0$ and $\rho = 1$ have relatively good forecasting accuracies. The MdRAFE score horizon is oftentimes worse for different settings of R_0 than the RMSFE and MSFE accuracy measures.

How can an informative prior be defined on the basis of this information? It should first be noted that the accuracy image will change if the initial reference

model R_0 is altered. When $\rho = 1$ instead of $\rho = 2$ and the number of experts is five instead of twenty, for instance, it is clearer that a track record of one can lead to poor forecasting accuracy, and that for the remaining part, longer track records increase prediction errors. When looking at different initial reference models, forecasting horizons, and dependent variables, the general conclusion is that track records of eight or higher should receive smaller combination weights than track records between one and eight. Regarding the number of experts, it appears to be beneficial to use between four and twenty experts. When less than four experts are included, overall forecasting accuracy is often poor. Such general conclusion cannot be made about the score horizons. For $\text{NGDP}_{h=2}$, for instance, score horizons $rho = 0$ and 2 are best avoided, whereas for $\text{PGDP}_{h=2}$, $rho = 0$ and $rho = 1$ worked out best. Also for shrinkage rates we find that for some horizons forecasting accuracy increases as shrinkage rates get higher, and for other horizons they decrease. The optimal choice of score variables and evaluation functions also varies across horizons and variables. Hence, only for track records and number of experts an informative prior can be derived. For the other items, a flat prior is more appropriate.

Next to developing an informative prior, our automated algorithm can also be expanded, for instance, by increasing the number of perturbations assigned to score horizons. The reasons for focusing on score horizons are that score horizons have a large influence on forecasting accuracy, and that some score horizons consistently outperform others for given forecasting horizons and dependent variables. In the automated algorithm, AA.a models were generated based on a single score horizon and score horizons were combined after the best AA.a models were selected. Instead, an expert's (or model's) combination weight could be based on a weighted average of forecasting errors from different score horizons, where the length of the track record could be varied across score horizons as well. Another way that AA could be expanded is by including more score variables.

To conclude this subsection, even with the benefit of hindsight we cannot find a set of configurations that outperforms Mean SPF for all horizons of PGDP and NGDP. On the other hand, it is clear that the previous-best forecast and the inverse pMSFE method were improved by the algorithm. Poor results were often attained when track records were long, the number of top-ranked experts to be included was too small or too high, the score horizon was the same as the forecasting horizon, and when only a pMSFE accuracy measure was used. The automated algorithm can in turn be improved by defining an informative prior which avoids models with less than four experts and a track record length higher than seven. The algorithm can be expanded by increasing the size of the set of candidate score horizons and score variables when combining expert or model forecasts.

5.4 Restricting AA items

If an analyst would want to expand the number of configurations of score horizons when AA is used for some other application, then she also needs to think about how to restrict the number of perturbations in AA. This can be done by studying the average amount of variation in predictions across the configurations (S_i) of a particular item i, see equation (10) in subsection 4.4.

To begin with the importance of windows and starting points in AA.c and AA.b, Table 3 is presented. For AA.c, on average 68% of the predictions of different window sizes lie within the tiny interval of only .02 around the mean of all the windows at a given time for $\text{PGDP}_{h=2}$. Such small values for S are also found for other horizons of PGDP and NGDP. The choice of window for NGDP is more influential at step AA.b, since S is either fourteen or fifteen. In line with Figure 5, the mean variation across starting points is small for AA.c and large for AA.b.

Figure 7 shows how S is affected by changing one of the other restriction values in the reference model R_0. The darker a rectangle is, the larger is S.

Table 3: S across windows and starting points

Variable	Aspect	Model	$h=0$	$h=1$	$h=2$	$h=3$	$h=4$
PGDP	Windows	AA.c	.01	.01	.01	.03	.02
		AA.b	.08	.08	.07	.10	.10
		TCOMB-SIC-Bias	.03	.06	.08	.09	.09
	Starting points	AA.c	.03	.04	.02	.05	.06
		AA.b	.47	.35	.19	.19	.24
		TCOMB-SIC-Bias	.56	.57	.28	.34	.40
NGDP	Windows	AA.c	.03	.05	.03	.04	.01
		AA.b	.15	.15	.14	.15	.14
		TCOMB-SIC-Bias	.00	.07	.11	.14	.11
	Starting points	AA.c	.04	.10	.08	.09	.02
		AA.b	.77	.62	.27	.33	.34
		TCOMB-SIC-Bias	.84	.87	.27	.37	.33

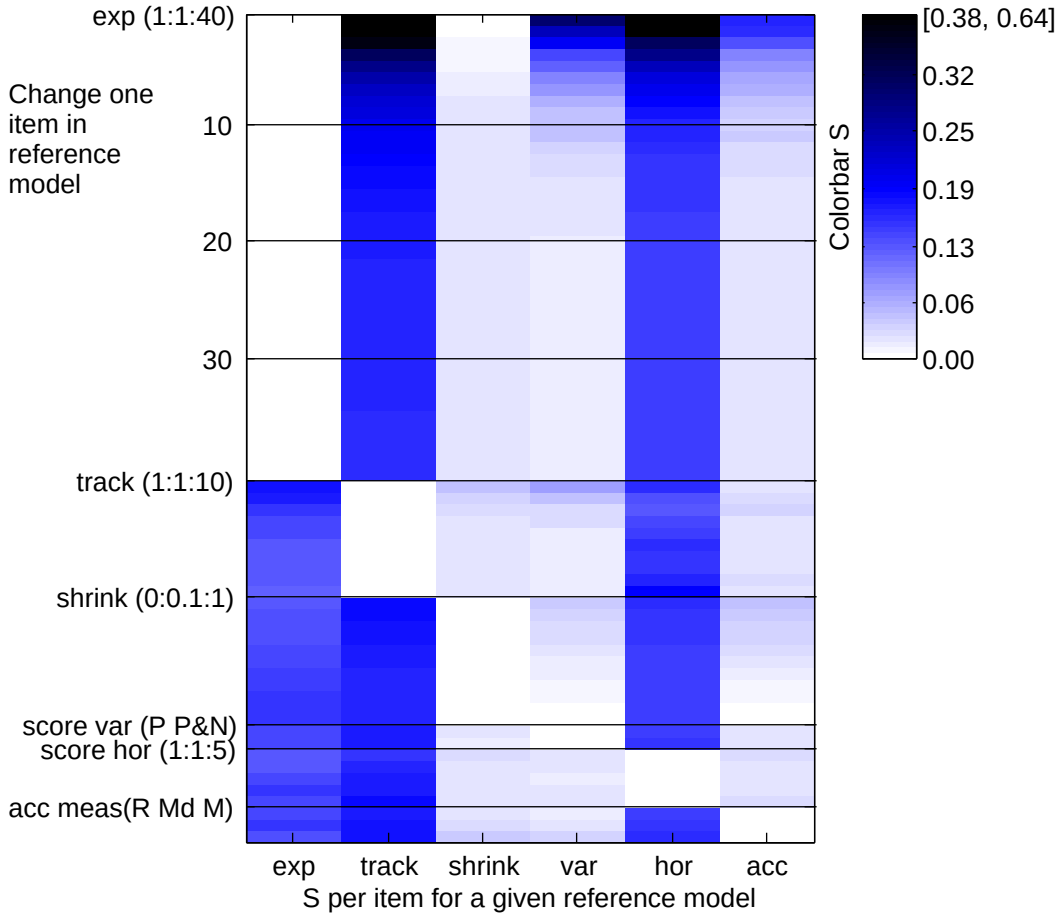
Gives S, the mean standard variation in forecasting errors, across different windows or starting points, for TCOMB-Bias, AA.c, and AA.b. The closer S is to zero, the smaller are the differences in forecasting errors across windows or starting points. To compare forecasting errors across different starting points, a window of twenty observations was used.

Table 4: Reference model R_0 and forecasting variations S for $\text{PGDP}_{h=2}$

Item	R_0	S_{R_0}	$\min S_R$	$\max S_R$
Experts	20	0.14	0.12	0.18
Track record	5	0.17	0.16	0.64
Shrinkage	1	0.03	0.00	0.06
Score variable	PGDP	0.03	0.01	0.30
Score horizon	2	0.15	0.13	0.57
Accuracy Measure	RMSFE	0.03	0.00	0.17
Experts (20:1:40)		0.00	0.00	0.02
Shrinkage (0:1:1)		0.06	0.00	0.12
Accuracy measure (RMSFE, MSFE)		0.02	0.00	0.05

This table shows for each item the standard reference model R_0, the mean standard variation S_{R_0} over all configurations of one item given R_0, the minimum S_R and the maximum S_R, as used in Figure 7.

Figure 7: Influence of changing a restriction for different reference models for $\text{PGDP}_{h=2}$



This Figure shows the mean standard variation (S) across all configurations of an item for a given reference model, whereby the reference model is altered on the vertical axis. The darker a rectangle, the larger the variations in forecasts are across different configurations of an item are. The rectangle in the first row and the third column is white. It represents zero variation because the shrinkage level does not change forecasts when only one expert is included. One column to the left shows that the average standard variation in forecasts when the track record is altered is .64 in case the reference model has one expert. This is the highest amount of variation in the figure, which is why the rectangle has a black color.

The benchmark values and the extreme values of S are shown in Table 4. The dark blue bar in the first column and the first row of 'Track (1:1:10)' represents $S_{\text{Experts}}(R^1_{\text{Track}}) = .18$; it is the average standard variation in prediction errors caused by changing the number of experts relative to a reference model which contains a track record of one instead of five. When the reference model is redefined in terms of the shrinkage rate, S_{Experts} gradually increases from .13 to .16. If each expert receives the same weight, then the choice on the number of poorly-ranked experts to exclude becomes more important. Looking at the first column, the general tendency seems to be that the decision on the number of experts becomes less important when the track record increases. In line with the expectations expressed in section 2, the decision on the track record becomes more influential as more experts are excluded and the degree of shrinkage towards the mean decreases. The same goes for the score variable, score horizon, and accuracy measure in the last three columns. As the third column shows, the overall influence on predictions of the shrinkage rate is small. The effect of shrinkage gets larger when different accuracy measures are used, because the RMSFE already shrinks weights towards the mean. Looking at the last two rows of the S image, the effects of changing the track record and the score horizon on the average variations in prediction errors of the other items are less straight-forward. The plots are similar for other horizons of PGDP and NGDP. As the third to last row of Table 4 indicates, there is very little difference in prediction when the number of experts is varied between twenty and forty with increments of one; the maximum $S_{\text{Experts (20:1:40)}}$ is only .02. The amount of variation in predictions across different shrinkage rates remains low when only zero shrinkage and full shrinkage are considered. As the last row of this table indicates, the difference in AA.a predictions between RMSFE based scores and MSFE based scores is also small (max $S_{\text{Acc meas (R, M)}} = .05$).

Now, to reduce the number of perturbations without affecting the forecasting accuracy of the automated algorithm, we would suggest to take the equal-

weighted mean of the three accuracy measures instead of applying TCOMB in step AA.c. Where the difference in accuracy measure is quite large when applying TCOMB in steps AA.c and AA.b, it is pretty small when expert forecasts are combined. So, for AA.a models only MdRAFE and MSFE could be pooled. For TCOMB in step AA.b, the set of window sizes can be defined as 10:5:30 instead of 10:1:30. The amount of candidate shrinkage rates used for combining expert forecasts can also be reduced considerably, for instance to $\phi = 0, 1$. The set about the number of top ranked experts can be brought down to (1,2,. . . ,20, 30, 40) without affecting the forecasting accuracy of the automated algorithm.

5.5 TCOMB-SIC-BIAS

Having discussed the way AA performed, the way AA updated configurations, the way AA can be improved, and the way AA can be restricted, we shall now continue with analyzing these four aspects for TCOMB-SIC-bias.

Table 5 shows the forecasting accuracy of SIC-bias, TCOMB-SIC-bias, and SIC-bias under the Pesaran and Timmermann setup. PGDP is a variable where SIC-bias performed particularly well in Cápistran and Timmermann. When SIC-bias is used over the entire sample, instead of with 1981Q3 as a starting point, the Mean SPF forecast is always chosen by the SIC criterion for both PGDP and NGDP, except for some predictions of $\text{NGDP}_{h=0}$. Since the RMSFE results are divided by the RMSFEs of Mean SPF, the scores for SIC-bias are therefore nearly always one. If one optimizes over the starting point of the sample using TCOMB with an MSFE evaluation function ('TCOMB-SIC-bias'), the bias-corrected model is preferred by the SIC criterion more often. As Table 5 shows, TCOMB-SIC-bias appears to perform slightly better than SIC-Bias for PGDP and slightly worse for NGDP. Given a window of size 10, it is better to take a weighted average of all starting points than to select the single best starting point for NGDP, and for PGDP it is the other way around. The results of TCOMB are comparable to taking the single best starting point with

Table 5: SIC-bias results relative to Mean SPF (1968Q4:2011Q3)

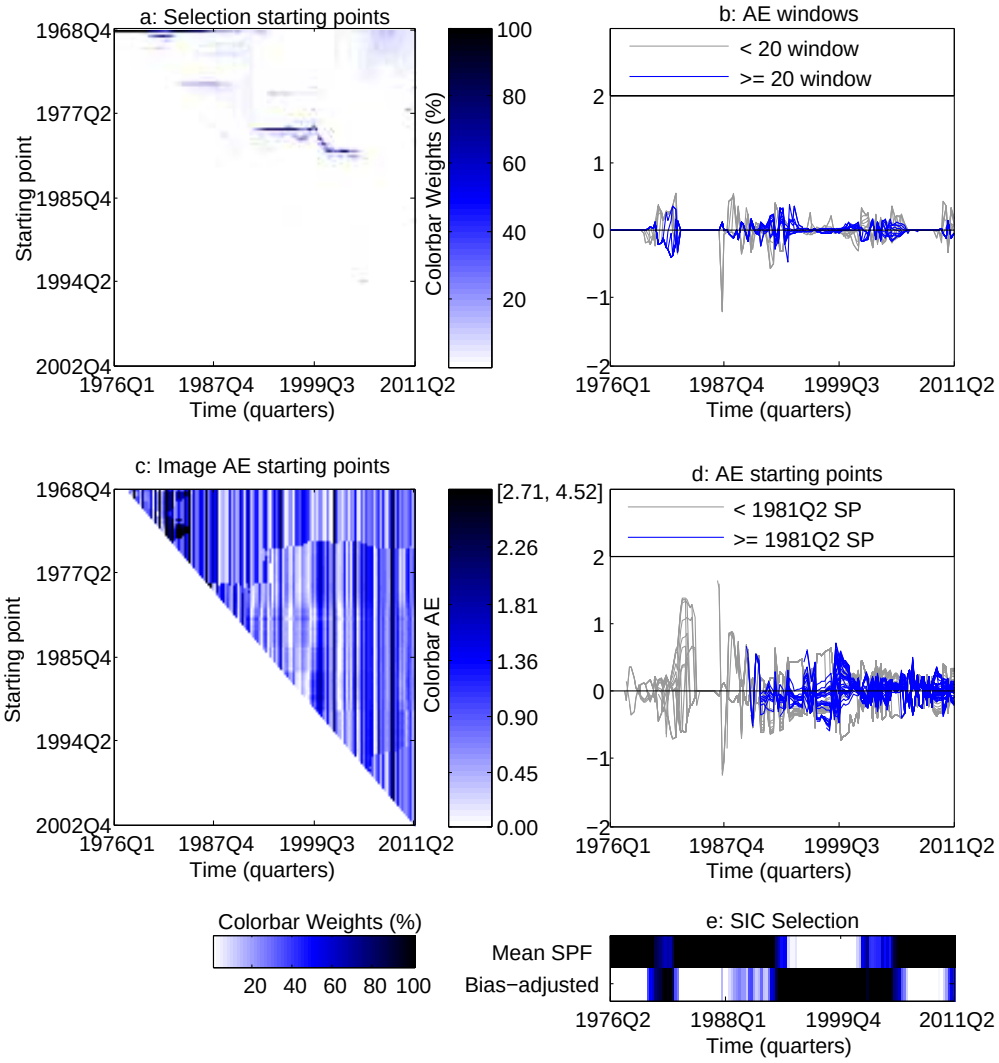
Variable	Evaluation	Model	$h=0$	$h=1$	$h=2$	$h=3$	$h=4$
PGDP	RMSFE	SIC-bias	1.00	1.00	1.00	1.00	1.00
		T-COMB-SIC-bias	0.99	0.99	0.99	0.99	1.00
		PT-SIC-bias-one-W10	1.00	1.01	1.01	1.04	1.04
		PT-SIC-bias-one-W20	0.99	0.97	0.99	0.98	1.00
		PT-SIC-bias-all-W10	1.00	1.00	1.02	1.07	1.12
		PT-SIC-bias-all-W20	1.00	1.00	1.02	1.07	1.12
NGDP	RMSFE	SIC-bias	1.01	1.00	1.00	1.00	1.00
		T-COMB-SIC-bias	1.01	1.00	1.01	1.02	1.00
		PT-SIC-bias-one-W10	1.01	1.01	1.02	1.07	1.03
		PT-SIC-bias-one-W20	1.01	1.01	1.01	1.01	1.00
		PT-SIC-bias-all-W10	1.01	1.00	1.01	1.05	1.01
		PT-SIC-bias-all-W20	1.01	1.00	1.01	1.03	1.00

The RMSFE of the model is divided by the RMSFE of Mean SPF to obtain the results. The same goes for the other evaluation functions. So, when the result is smaller than one, the model outperformed Mean SPF; and vice versa. PT stands for the Pesaran and Timmermann setup for optimizing over the starting point of the treatment sample. The addition 'one' means that only the single best starting point was selected, and 'all' means that a weighted average of all available starting points was used. 'W10' means that a window of ten observations was used to evaluate starting points. The MdRAFE results were all 1.00 and have therefore been left out.

a window of twenty observations (PT-SIC-bias-one-W20).

Figure 8.a is a distribution image which shows the weights that were assigned to the starting points for $\text{PGDP}_{h=2}$. A blue dot means that a positive weight was given to that starting point. The darker the spot, the higher the weight. When a spot is completely white, the forecasts based on that starting point were not included. The first moment that a choice between candidate starting points is made is in 1977Q4. Before that time, the earliest starting point of the treatment sample is always used. As of 1992Q2, starting points of around 1981Q2 become popular. At that point in time, SIC selected the bias-adjusted model, as can be seen in Figure 8.e. Due to some poor bias-adjusted forecasts, the full sample is used once again in 2005Q4, causing the SIC criterion to choose for Mean SPF. The distribution image would contain one black rectangle for each column in case the PT setup was used whereby only the best starting point is selected. If a distribution image of the starting points were made for PT-SIC-

Figure 8: Starting points and absolute errors TCOMB-SIC bias for $\text{PGDP}_{h=2}$



a. This panel shows the weights assigned to the starting points (vertical axis) over time (horizontal axis). As the colorbar next to the graph indicates, the darker a dot, the higher is the weight attributed to that starting point.
b. This panel shows the absolute error of a particular window size minus the absolute error of taking the mean of all the window sizes. When this difference is negative, that particular window size outperformed the average forecast of all the window sizes.
c. This panel shows an image of the absolute errors of particular starting points. The darker a rectangle is in a given row, the larger the forecasting error was for that particular starting point. If a column has the same color, this means that there was no difference in forecasting accuracy across starting points.
d. This panel shows the absolute error of a particular starting point minus the absolute error of taking the mean of all the starting points for a window of evaluating the starting points of size twenty. When this difference is negative, that particular starting point outperformed the average forecast of all the starting points.
e. This panel shows when the SIC criterion chose Mean SPF and when it selected the bias-adjusted model.

bias-all, then all the available starting points would have had some light shade of blue. By contrast, in TCOMB, only a number of best-ranked starting points are combined, leaving out starting-points that harm forecasting accuracy. This explains the difference in forecasting accuracy in Table 1.

To further illustrate this point, consider Figure 8.c, which is an image of the absolute errors made by different starting points (vertical axis) at a given time (horizontal axis) for $\text{PGDP}_{h=2}$. When a part of a column has the same color, there are no differences in absolute errors across starting points for that part. This way it can be seen that the absolute errors (AE) for starting points between 1968Q4 and 1973Q4 are nearly always the same. When a column is darker for some starting points than for others, then these starting points had higher absolute forecasting errors. It can thus be observed that starting points between 1968Q4 and 1973Q4 are often worse (darker) in the period between 1983Q2 and 1996Q3 than starting points around 1981Q2. Hence, it would be better to leave out the observations between 1968Q4 and 1973Q4 in estimating SIC-Bias between 1983Q2 and 1996Q3. In the final part of the sample, the starting points before 1973Q5 generally have lower AEs, making it attractive to incorporate them once more. In Figure 8.a., the blue blur at the end of the sample for starting points before 1973Q5 which is explained by the fact that the AEs for these starting points are all so similar.

To examine whether the choice of window size used for comparing different starting points is important, Figure 8.b is presented. Absolute errors of window sizes are compared to the mean absolute forecast errors of all windows. The only difference between windows that are smaller than twenty and windows that are larger appears to be that the forecasting accuracy of smaller windows are more volatile for $\text{PGDP}_{h=2}$. The average variation in forecasting accuracy for different window sizes is small. As Figure 8.d indicates, larger variations in forecasting accuracy are found when starting points are varied. The average standard variation in forecasting accuracy across different (S) window sizes at

a given time is presented in Table 3. For $h = 2$, around 68% of the forecasting errors of different windows at a given time lie within .12 of the mean. If one would compute the standard variation in AEs across starting points at each time in 8.d, and would take the mean of all those standard variations, then one would find that, when the starting points are varied, the mean standard variation at a given time is a lot more, namely S=.28. Hence, an *ad hoc* choice about the size of the window used for evaluating the starting points has less influence on final forecasts, whereas the *ad hoc* choice about the starting point of the sample is more important.

To sum up the results of SIC-bias, we conclude that the choice of the starting point appeared to be pivotal for the results. Therefore, it might be worthwhile to add configurations in the algorithm for starting points. For example, one might optimize over the treatment sample of the bias-adjusted model first, and optimize over the treatment sample once again to let the SIC-criterion decide between this TCOMB-bias-adjusted model and Mean SPF. To speed up computing time, fewer window sizes can be included for evaluating the performance of different starting points.

5.6 Conclusion

Based on the analysis of AA and TCOMB-SIC-bias, we have robustified the claim that the equal-weighted forecast is difficult to beat. We further optimized over certain *ad hoc* judgments concerning the length of the track record, the number of forecasters included, the degree of shrinkage towards the mean forecast, the score variables and score horizons used in computing individual scores, the starting point of the treatment sample, and the accuracy measures used for combining and evaluating forecasts, and we still could not beat Mean SPF. It has been shown how the algorithms updated the selection of configurations and we have studied how the algorithms might be improved by incorporating an informative prior and expanding the number of configurations for

certain items. We have also shown which items could receive less configurations without thereby affecting final predictions. It remains doubtful whether these adjustment will sufficiently improve AA and TCOMB-SIC-bias to enable them to beat the Mean SPF.

Many parts of AA can be useful for other applications of time series forecasting. Although the algorithms could not beat Mean SPF, they certainly improved the models of CT on which they were based; namely, SIC-bias, previous-best expert forecast, inverse pMSFE method, and the shrinkage model. The improvements resulted from a considerable expansion of the set of candidate configurations and from the way that configurations were selected and combined. Rather than selecting the single best configuration or all of the candidate configurations, COMB combined a number of best ranked configurations. In this way it was shown, for example, that it is better to combine between four and twenty experts instead of just the single-best expert forecast or all of the eligible expert forecasts; that long track records should be avoided; and that poor performing score horizons and starting points can at times best be ignored altogether.

6 Conclusion

In this paper we have advocated the use of automated algorithms in order to robustify empirical econometric analysis. With a detailed illustration we have shown the three benefits of employing automated algorithms to update *ad hoc* decisions. First, we documented the uncertainty involved in the decisions that have to be made in real time by updating the analyst's *a priori* conjectures using the available data. Second, by optimizing over pseudo out-of-sample observations, smaller and more consistent prediction errors could be achieved in comparison to a non-robust setting. Third, the application of automated algorithms reduces data-snooping, whereby a researcher chooses *a posteriori*

what the best set of configurations is, such as the choice between a rolling window and an expanding window (Pesaran and Timmermann, 2007).

In our paper we expanded on CT (2009), who have shown that in spite of their well appreciated efforts, the mean survey forecast is extraordinarily difficult to beat. To further robustify this stylized fact, we began with identifying some of their *ad hoc* decisions, which are decisions that are generally made in such a way in time series forecasting. First, in giving weights to experts, CT either selected the best ranked forecaster or they selected all the forecasters. Second, forecasters received weights if they had a 'sufficiently long' track record. Third, the shrinkage factor was determined to be either 0.25 or 1. Fourth, when combining forecasts of one particular variable, only the forecast errors of that particular variable were used to give combination weights to experts. Fifth, when combining h-period-ahead forecasts, $\rho \neq h$ period ahead forecast errors were not considered when computing weights to pool forecasts. Sixth, the starting point of the expanding window was always 1981Q3. Finally, model forecasts were evaluated by an RMSFE accuracy measure while expert forecasts were pooled based on inverse pMSFE scores.

We developed and illustrated an automated algorithm to show the degree of uncertainty of these *ad hoc* decisions and to make the selection of configurations more dependent on data. Our AA algorithm was based on the well known idea of creating pseudo out-of-sample forecasts for varying configurations and selecting the model which resulted in the smallest forecasting errors. A sub-algorithm called 'COMB' combined a number of best-ranked models with shrunken inverse score weights, and another algorithm 'TCOMB' also optimized over the starting point of the treatment sample when pooling forecasts. TCOMB was also applied to a method of CT whereby a SIC criterion was used to decide between a bias-adjusted model and the mean survey forecast.

In the analysis of our results, we presented distribution images which displayed the relative frequencies with which configurations were selected in order

to learn how preceding estimates were updated over time. We found that, in order to optimize forecasts in terms of one particular evaluation function, it was useful to consider other evaluation functions in the algorithm. In a similar vein, $(\rho \neq h)$-period-ahead forecasting errors turned out to be helpful when computing weights that were used for combining h-period-ahead forecasts at the next stage. Another interesting finding was that starting points after the structural break of the seventies of the twentieth century were often selected once they became available. And, in the last part of the sample, observations at the start of the sample were sometimes reintroduced to capture some of the turbulent dynamics of the past. A remarkable fact that has hitherto been unobserved to our knowledge was also illustrated; namely, that different items benefit from having different starting points of the treatment sample. The number of forecasters, the length of the track record, and the degree of shrinkage towards the mean varied across starting points, score horizons, forecasting horizons, accuracy measures, variables, and time.

With the help of an accuracy image we studied whether the algorithms improved the models on which they were based and whether the algorithms themselves could be enriched. It was shown that *ad hoc* choices can have substantial effects on forecasting accuracy. Where CT either selected the single best expert or all of the experts with a sufficiently long track record, we have shown that it is better to use between four and twenty experts. It also became clear that long track records should be avoided. AA can be improved by developing an informative prior based on this information and by expanding the number of configurations concerning the score horizons and score variables. To our knowledge, score horizons other than the forecasting horizon and score variables other than the dependent variable have so far not been used for selecting and combining models. Particularly when data is scarce, it will be beneficial to consider different score horizons and score variables. It was also shown that TCOMB outperformed the *ad hoc* decision of using the first starting point and

that TCOMB outperformed the Pesaran and Timmermann setup whereby a weighted average was taken over all of the available starting points. Other methods for selecting and combining optimal configurations could be studied.[7]

By plotting or tabulating the average standard variation in predictions across different configurations of a particular item, it was shown that many configurations can safely be ignored without affecting final predictions. As it turned out, the number of windows used for evaluating different starting points can be reduced considerably. The shrinkage rate is not so important when combining top-ranked experts. Also, there is barely any difference between pooling forty or twenty of the top ranked experts, which implies that the configuration space of the number of experts can be reduced considerably.

Of course, our study is not the definitive one on this topic. Indeed, the irony of this paper is that to update and evaluate CT's *ad hoc* decisions, we had to introduce our own new *ad hoc* decisions. For example, to confront the decision on the starting point of the treatment sample with data, we needed to decide on the size of the moving window used for evaluating the performance of different starting points and this decision was not updated by data. Other such *ad hoc* decisions are the following. We evaluated forecasters based on flash realizations, we ignored density forecasts, we only looked at combinations between top-ranked forecasters, we only used a single set of AA.a configurations per score horizon, and there are many more.

Will there ever be an end to updating *ad hoc* decisions? We think not. But, with the help of the tools presented in this paper, we can identify which configurations can safely be ignored, and which configurations should be incorporated in the algorithm in order to optimize forecasts. In this way, automated algorithms can instigate new ideas which could further robustify or overthrow old ones. Accordingly, by refining some of the methods of CT with an automated algorithm, we have strengthened their claim that an equal-weighted

[7] Machine learning literature, see for instance Hastie et al. (2009).

mean should be used when pooling experts' forecasts.

A Programming Code

This document accompanies the essay 'Some Tools for Robustifying Econometric Analyses.' A concise explanation is presented for the Matlab code that was used in this paper. Table A.6 gives an overview of the matlab functions used in the paper. To help the reader to get to grips with the program, we will further discuss the most important functions. The programs ('functions') used for generating the figures and tables in the paper are summarized in Table A.6.

The function *index* contains all the functions required for making ($h = 0, 1, \ldots, 4$)-period-ahead-forecasts for one SPF variable with our automated algorithm AA and Time SIC Bias. The reader can just run 'indexAA(PGDP, IpredPGDP, NGDP, IpredNGDP,indPGDP)' in the command window to generate all the forecasts, tables, and figures.

A.1 Preparation

The function *prepare* runs the functions *realdata*, *dat*, *qfind*, *SPFmed*, and *ErInd*, in order to transform the realizations and individual predictions into the variable ***DV*** and ***Qtot*** respectively. The output of the function *realdata* is the dependent variable ***DV***, it is shown in Figure 9. The first column gives the time indication (1970, 1970.25, 1970.5, 1970,75 correspond to 1970Q1, 1970Q2, 1970Q3, 1970Q4 respectively). The second column are the realizations, based on flash observations. The scalar 3.8413 in the first row and the second column corresponding to inflation in 1968Q4 is calculated by $400 * \ln \frac{PGDP_{1968Q4}}{PGDP_{1968Q3}} = 400 * \ln \frac{123.46}{122.28}$.

The function *qfind* generates ***Qtot***, a matrix which contains the individual forecasts, based on input ***Ind*** given by the function *dat*. Figure 10 shows what the matrix looks like. The expert forecasts are aligned next to each other and each expert has seven columns, so that a 172 by 2982 matrix is produced. The first column is the expert ID, the second the time indication, and the subsequent

five columns are the $(h = 0, 1, \ldots, 4)$-period-ahead-forecasts. In 1968Q4 the first forecasts are made. The number 3.2653 in row one (1968Q4) and column four (h=0) is calculated by $400 * \ln \frac{y^1_{1968Q4|t=1968Q4}}{y^1_{1968Q3|t=1968Q4}} = 400 * \ln \frac{123}{122}$; it is the $(h = 0)$ period ahead forecast made in 1968Q4 by expert with id 1. The scalar 3.2389 in row two and column four is the $(h = 1)$ period ahead forecast made in 1968Q4 by expert with ID 1. All $h = 0$ to $h = 4$-period-ahead-forecasts made about $400 * \ln \frac{y_{1969Q1}}{y_{1968Q4}}$ are shown in the second row of ***Qtot***. The first expert did not submit any predictions in 1969Q1, which explains the NaNs running diagonally from (2,3) to (5,6). The seven columns of the second expert start in column eight. This format is also convenient for analyzing performances of individual experts.

The function *ErInd* gives the forecast errors in the format of ***Qtot***. The individual forecast error for the $(h = 0)$ period ahead forecast made in 1968Q4 by expert 1, is given by 3.2653-3.8413=-0.5760. This number squared is shown in ***sFE***, and the absolute version in ***aFE***. The relevant output of the function *SPFmed* are the ***meanSPF*** forecasts. Columns 1 to 5 give the $h = 0$ to $h = 4$-period-ahead-forecasts. The function *scoreRAE* gives the absolute errors relative to the absolute errors of ***meanSPF*** for each forecaster in the ***Qtot*** format. The number on row 1 column 3 is 1.1751 and it follows from dividing the absolute forecast error of expert 1 by the absolute error of the mean SPF (0.5760/0.4902). Apparently, this forecast was worse than the average forecast.

A.2 AA.a

In AA.a the experts forecasts are combined by varying the score horizons, the shrinkage rate, the size of the track record, and the number of best-ranked experts pooled. The functions in this section are (in)directly part of the function *AAa*.

The function *ranka* ranks the experts. The input variable ***W*** contains individual SEs or RAEs (depending on ***evalu*** in *optpi*) in a ***Qtot*** format.

The scalar ***sch*** indicates the score horizon, ***numlag*** refers to the the required length of the track record and ***nc*** sets the maximum amount of experts included. The score variable is indicated by ***twoDV***; if it is one, only PGDP is used for predicting PGDP, and if it is two, both PGDP and NGDP forecasting errors are used for assigning a score to experts. In the first part of *ranka*, experts receive an RMSFE, MRAFE, or MSFE score, based on the required number of recent *available* observations and the score variable. Experts are subsequently ranked. In the final part of *ranka*, the scores that are computed using $\rho-$period ahead forecasts errors are assigned to $h-$period ahead expert forecasts. The structure T has matrices ***Tax***, ***Taq***, and ***indIXa***. The rows of these matrices correspond to time. The first column of ***Tax*** contains the score of the best ranked expert, the second column the score of the second-best ranked expert, etc. Similarly, the first column of ***Taq*** contains the forecast of the best ranked expert, the second column the forecast of the second-best ranked expert. ***indIXa*** shows the ranking of the experts as indicated by their id. The letters $a, \ldots, e$ in ***Taq***, ***Tbq***, etc., refer to $h = 0$ to $h = 4$ period ahead forecast.

Let us generate some output using $[T] = ranka(W, Qtot, 1, 1, 40, 1, 1)$, where W are SEs. These specifications mean that the single most recently available ($\rho = 0$) period ahead forecast error was used to give a score to the experts; a maximum of 40 experts are included; only the dependent variable is used as a score variable; and an RMSFE accuracy measure is used (last input). Say we are interested in ($h = 1$)-period-ahead-forecasts, so that we look at ***indIXb***, ***Tbq***, and ***Tbx***. The first ten rows and seven columns of each variable are displayed in Figure 11. The first time that an ($h = 1$)-period-ahead-forecast of the expert with ID one is selected is for the $400 * \ln \frac{y^1_{1969Q4|1969Q3}}{y^1_{1969Q3|1969Q3}}$ in row five, column three. The ($\rho = 0$)-quarter ahead-forecast-error about 1969Q2 is known in 1969Q3 and used to select the expert's ($h = 1$)-quarter-ahead-forecast made in 1969Q3 about 1969Q4 (row 5, column 2 in ***Qtot***). The reader will have noticed that expert number 33 made a forecast of zero (row 5, column 2) and might find this

to be a remarkable outlier. The number zero arises when the expert forecast of the untransformed PGDP is the same for two subsequent periods. This is the case for expert 33, whose untransformed nowcast and one-period forecast in 1969Q3 were 128 and 128.

The function *optpi* applies *ranka* for different score horizons, score variables, and sizes of the track record. $\boldsymbol{S}$ is a (10×5) structure, as there are ten different sizes of the track record and five different score horizons. Within an element of $\boldsymbol{S}$ are $\boldsymbol{indIXa}$, $\boldsymbol{Tba}$, and $\boldsymbol{Tax}$, and so on, again.

In *optmedShrink* the expert forecasts of one column of $\boldsymbol{S}$ are combined using inverse shrunken scores as weights. The shrinkage factor is given by $\boldsymbol{phi}$ and the maximum number of $\boldsymbol{S}$ ($\rho = 0$), zero shrinkage, and a maximum of forty experts, we run $[MaW, MbW, McW, MdW, MeW] = optmedShrink(S(:, 1), 0, 40)$. The output, $\boldsymbol{MaW}$, $\boldsymbol{MbW}$, etc., are (172×400) matrices, with time on the vertical axis and forecasts of different perturbations on the horizontal axis. 400 perturbations arise out of the 10 different lengths of the track record and the maximum of 40 top-ranked experts which are at most combined. In the function AAa they are stored in a structure called $\boldsymbol{yAAa}$. Figure 12 shows the first ten rows and seven columns. It will be noticed that the first column of 12 is the same as the first column of 11, because the first column contains the single best ranked experts. In the third column, a weighted average of three of the best ranked experts is taken. Let us check the value of 2.0219 in row five column three. The inverse score weights of 2.2838, 2.2838, and 2.7253 (see $\boldsymbol{T.Tbq}$) are 0.3524, 0.3524, and 0.2953. When these weights are shrunken towards the mean with a factor of 0.1, they become 0.3505, 0.3505, and 0.2991. Taking the inner product with the forecasts 3.1129, 0, and 3.1129 gives the pooled forecast of 2.0219.

The function *comboptmed* applies *optmedShrink* for different score horizons and shrinkage rates. The output are $\boldsymbol{Ma}$, $\boldsymbol{Mb}$, etc., which are $(172 \times 22,000)$ matrices. Again, the letters $a, b, \ldots, e$ refer to $h = 0$ to $h = 4$-quarter-ahead-

forecasts. The 44,000 perturbations come from five score horizons, two score variables, eleven shrinkage rates, ten track record sizes, and forty expert combinations. The function *errorM* gives the prediction errors using squared errors or relative absolute errors of all these forecasts.

A.3 TCOMB

TCOMB combines models and optimizes over the starting point by assigning shrunken inverse score weights to a number of top-ranked models/starting points. Section A.5 shows simulation results of TCOMB. In AA.b *TCOMB* is used to combine the best AA.a models of each score horizon with meanSPF for each accuracy measure, so that we end up with three sets of forecasts (RMSFE, MdRAFE, and MSFE). *TCOMB* is made up of *COMB*, *TCOMBi*, *TCOMBii* and *TCOMBiii*.

The function *COMB* combines point forecasts of various models by assigning shrunken inverse score weights to a number of best-ranked models. If the treatment sample runs from $\boldsymbol{q}(1)$ to $\boldsymbol{q}(2)$, then the input variables ***ypredtr*** (pseudo out-of-sample model forecasts), ***DV*** (realizations), and ***SPFr*** (truncated absolute meanSPF errors), should contain $\boldsymbol{q}(1)$ to $(\boldsymbol{q}(2) + 1 + h)$ data points. The outcomes of the treatment sample are known at $(\boldsymbol{q}(2) + 1)$, at which point an h period ahead forecast is made. So, a real-time $(h = 0)$-period ahead forecast is made about $(\boldsymbol{q}(2) + 1)$. In the program, $j = h + 1$ is an input variable. Models that contain missing observations are deleted. Models are subsequently ranked based on their scores and a number of best ranked models are combined with varying shrinkage rates. Scores are again computed for the combined models, and the set of models with the lowest score is selected. The output variable ***indSel*** saves the shrinkage rate (column 1), the k selected models (column 2 to $k + 1$) and the weights assigned to each model (column $k + 2$ to $2k + 1$). ***ONEpred*** gives a single prediction point. The last output ***bench*** is a row vector containing the single prediction point in the first col-

umn, the Pesaran and Timmermann prediction in the second column (combine all available starting points using an inverse pMSFE evaluation function and a window of twenty observations to evaluate starting points), and the predictions of all the models in the remaining columns.

In *TCOMBi*, the forecasts of models are generated for different starting points of the treatment sample. When ***indM***=1, *TCOMBi* runs *combAAa* for each starting point of the treatment sample and saves the forecasts and specifications in ***startM***. The latter program selects the best forecast of each score horizon and combines these forecasts with meanSPF using *COMB* in an expanding window setup with a prespecified starting point. The input variables are ***start***, which specifies the starting point of the treatment sample; ***DV***, all the realizations; ***Qtot***, individual forecasts; ***yAAa***, a structure containing the $(h = 0)$ to $(h = 4)$-period-ahead forecasts of all the AA.a models; and ***evalu***, which indicates which function is used for evaluating the forecasts. The output variable ***num*** specifies which of the 44,000 perturbations were selected; it is a (171×25) matrix. The first five columns of ***num*** contain the selected models of each score horizon for the $(h = 0)$ forecast, and the next five columns are the selected models of $(h = 1)$, etc. The other output variables are ***yAAb*** and ***RES***. The (171×5) matrix ***yAAb*** contains all the resulting forecasts for each horizon. ***RES*** is a structure which saves ***indSel*** of *COMB*. For each forecast, it is shown which models were included (score horizons and meanSPF), the shrinkage rate, and the weights assigned to each model.

TCOMBii uses the structure ***startM*** produced in *TCOMBi* as input. It applies *COMB* to pool the starting-point dependent forecasts in a rolling window setup. One of the input variables is *add*, which determines the size of the moving window used for evaluating the starting points. If there are not enough pseudo-out-of-sample starting points to compare starting points, a backup model is used. The output variable ***yii*** gives the pooled forecast of the five horizons, ***REST*** is a structure with the ***indSel*** specifications of *COMB*, and ***BENCH***

is a structure containing ***bench*** of COMB.

TCOMBiii applies *TCOMBii* for different sizes of the moving window (10 : 1 : 35) and takes the equal-weighted average of all the window-size dependent forecasts. The output variable ***Tres*** is a (21×5) structure containing the output ***REST*** for each of the twenty-one moving window sizes and each horizons. Element (1,3) contains ***rest*** of a moving window of ten and a horizon $h = 2$, for example.

In AA.c the final forecasts of each accuracy measure are combined using the algorithm called *TCOMB*, whereby *combAAb* instead of *combAAa* is used in *TCOMBi* to generate starting-point dependent combined forecasts (***indM***=2). The function *combAAb* applies *COMB* to combine forecasts in an expanding window setup for a given starting point.

The SIC-Bias model is calculated using *biasadjSPF* in *TCOMBi* (***indM***=3). The input variable ***start*** indicates the starting point of the sample. Output ***d*** shows whether SIC chose the bias adjusted model or not and ***SICbias*** contain the $h = 0$ to $h = 4$ forecasts of the SIC-Bias model for a given starting point.

A.4 Functions: Figures and tables

Having discussed the programming of the automated algorithm, we shall now continue to go over the programs used for generating output. An overview of the programs can be found in Table A.6. The function *tableRES* prints the results in terms of MSFE, RMSFE, and MdRAFE in a latex format. We have indicated which programs refer to which figures in the paper. Three programs, which are used in several plots, are further dicussed, along with the *plotsel* functions. The program *retrieveMOD* shows which AA.a restrictions are associated with a given row number of ***Ma***,..., or ***Me***. The displayed output (if ***disp***=*1*) for $retrieveMOD(10970, yAAa1.Ma, diso)$ is for instance give by

Score horizon	Shrinkage	# score variables	Track record	Experts
3	0.5	2	5	10

The matrix $\boldsymbol{R}$ consists of all possible permutations of AA.a restrictions, whereby the row of $\boldsymbol{R}$ corresponds to the column of AA.a. It can be added as an input variable ($\boldsymbol{Rg}$) to save computation time. To retrieve the column number of $\boldsymbol{Ma}$ of a given set of restrictions, *retrieveNUM* can be used.

For most plots, an explanation of the function has already been given in the paper itself. We can be more specific about all the plotsel function, which all work in the same way. The most 'complex' one is plotselAAa. Basically, we determine for a given evaluation function which windows were selected, which starting points were selected of each window, which score horizons were selected of each starting point, and which AA.a items were selected of each score horizon,

to compute the weights. Algorithm 4 shows how it is done.

Algorithm 4: plotsel-Algorithm

input : Tres, startAAb,h,cmap

output: Plots

for *t=1968Q4:Q:2011Q2* **do**

  for *Window = 10:1:total number of eligible windows (N_{win})* **do**

    Using Tres, find which starting points (SP^*) were used for combining score horizons

    for SP^* = *earliest:latest selected starting point* **do**

      Using startAAb, find which of the score horizons (sch^*) and Mean SPF were selected.

      for sch^*=*lowest:highest selected score horizon* **do**

        Find out which AA.a configurations were selected.

        Update weights of AA.a items, e.g. when e^* experts were selected: $W^{new}_{experts}(t, e^*) = W^{old}_{experts}(t, e^*) + 1 \cdot W_{sch}(sch^*) \cdot W_{SP}(SP^*) \cdot 1/N_{win}$

      end

    end

  end

end

Make images based on $W_{experts}$, $W_{shrinkage}$, etc.

A.5 Simulation TCOMB

To give an indication that *COMB* and *TCOMB* do what they are supposed to do, we shall perform a small simulation exercise and represent output using the functions in A.6.

$$y_t = \begin{cases} 0.4X_{1,t} + 0.6X_{2,t} + \epsilon_t & \text{if 1968Q4} \leq t \leq \text{1986Q1} \quad \text{(70 obs)} \\ 0.4X_{1,t} + 0.3X_{2,t} + 0.3X_{3,t} + \epsilon_t & \text{if 1986Q2} \leq t \leq \text{2011Q3} \quad \text{(102 obs)}, \end{cases}$$

where $X \sim$ N(0,1) and $\epsilon \sim$ N(0,0.25). From the model it is clear that there are 172 observations, of which seventy are generated by just two independent variables and the subsequent 102 observations by all three regressors. We shall enter the three regressors in TCOMB as if they are three forecasts. The algorithm of AA.c will be used, so that the minimum starting point is at the eleventh observation. What we should find is that the starting point of the treatment sample is past 1985Q4 once those starting points are available, and that the weights assigned to the variables should go from $\{.4, .6, 0\}$ to $\{.4, .3, .3\}$ after at least seventy observations.

Figure 13 shows the generated series, the selected starting points and the weights assigned to each variable. Until 1979Q4, the lowest available starting points is generally selected. After that time the most recently available starting point is used, which results in a kind of rolling window. The rolling window behavior is caused by the fact that data generated from a different process are included in the evaluation sample of the starting points. Once 1986Q2 is reached, that date is used as a starting point until 1989Q3 becomes available, which apparently has lower RMSFEs than the earlier starting points. The weights assigned to each variable are shown in the plot at the bottom of Figure 13. The average weights until 1990Q1 are $\{.43, .57, 0\}$ and in the second part of the sample they converge to $\{.26, .23, .51\}$.

This shrinkage towards the mean is caused by taking the root of the mean squared forecast errors. Figure 14 shows the shrinkage rate (column 1), the selected best-ranked variables (column 2 to 4), and the weights assigned to the variables (columns 5 to 7). Note that the shrinkage rate is always zero. By contrast, the plot below shows that for the MSFE evaluation function, the degree of shrinkage is around .3 and .4. These shrinkage rates cause the inverse pMSFE weights to converge to the actual parameters. Figure **??** shows what happens when MSFEs are used instead of RMSFEs. The selection of starting points displays quite a similar pattern, but the weights are closer to the actual

ones.

A.6 Appendix Figures and Tables

Figure 9: ***DV***

DV <172x2 double>

	1	2
1	1968.8	3.8413
2	1969	4.2629
3	1969.3	4.8381
4	1969.5	5.4670
5	1969.8	4.4010
6	1970	5.0226
7	1970.3	4.1185
8	1970.5	4.3340
9	1970.8	5.5544
10	1971	5.1272

Figure 10: ***Qtot***

Stack: Base | No valid plots for: Qtot(2,2)

Qtot <172x2975 double>

	1	2	3	4	5	6	7	8	9	10
1	1	1968.8	3.2653	NaN	NaN	NaN	NaN	2	1968.8	3.2653
2	1	1969	NaN	3.2389	NaN	NaN	NaN	2	1969	3.2129
3	1	1969.3	3.1873	NaN	3.2129	NaN	NaN	2	1969.3	NaN
4	1	1969.5	3.1373	3.1621	NaN	0	NaN	2	1969.5	NaN
5	1	1969.8	6.1540	3.1129	3.1373	NaN	3.1873	2	1969.8	NaN
6	1	1970	3.0418	3.0418	3.0888	3.1129	NaN	2	1970	NaN
7	1	1970.3	6.0152	3.0189	6.0152	3.0651	NaN	2	1970.3	NaN
8	1	1970.5	2.9740	2.9740	5.9703	2.9740	NaN	2	1970.5	NaN
9	1	1970.8	4.6956	2.9520	5.8825	2.9520	2.9520	2	1970.8	NaN
10	1	1971	2.6115	2.9070	5.8395	2.9091	NaN	2	1971	NaN

Figure 11: ***The structure T***

T.indIXb

T.indIXb <172x40 double>

	1	2	3	4	5	6	7
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2	4	7	8	9	10	15
4	4	16	23	26	31	32	43
5	26	33	1	3	6	10	15
6	3	14	23	40	46	47	49
7	3	13	14	15	16	31	40
8	55	69	75	103	15	32	40
9	6	8	22	23	32	38	48
10	1	7	15	16	18	21	23

T.Tbq

T.Tbq <172x40 double>

	1	2	3	4	5	6	7
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3.1873	3.2129	3.1873	3.1873	3.1873	3.1873	3.1373
4	3.1621	3.1621	3.1873	3.1373	3.1873	3.1873	3.1873
5	3.1129	0	3.1129	3.0888	3.1129	3.0888	6.2017
6	6.1070	6.1070	6.1070	3.0651	6.1070	3.0418	3.0418
7	3.0189	5.9703	3.0189	3.0189	3.0189	3.0189	3.0189
8	5.9260	2.9963	5.9260	2.9740	2.9740	2.9963	2.9740
9	2.9520	2.9520	2.9304	2.9520	0	5.8825	2.9520
10	2.9070	3.7832	3.4935	2.9091	2.9091	3.4935	4.0757

T.Tbx

T.Tbx <172x40 double>

	1	2	3	4	5	6	7
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	0.3318	0.3318	0.3318	0.3318	0.3318	0.3318	0.3318
4	1.0487	1.0487	1.0487	1.0487	1.0487	1.0487	1.0487
5	2.2838	2.2838	2.7253	2.7253	2.7253	2.7253	2.7253
6	0.6132	0.6132	0.6132	0.6132	0.6132	0.6132	0.6132
7	1.7219	1.7219	1.7219	1.7219	1.7219	1.7219	1.7219
8	1.0778	1.0778	1.0778	1.0778	1.1760	1.1760	1.1760
9	1.2092	1.2092	1.2092	1.2092	1.2092	1.2092	1.2092
10	1.8495	1.8495	1.8495	1.8495	1.8495	1.8495	1.8495

Figure 12: ***MbW***: combining $h = 1$ best-ranked expert forecasts using $\rho = 0$ forecast errors

MbW <172x400 double>

	1	2	3	4	5	6	7
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3.1873	3.2001	3.1958	3.1937	3.1924	3.1915	3.1838
4	3.1621	3.1621	3.1705	3.1622	3.1672	3.1705	3.1729
5	3.1129	1.5564	2.0219	2.2674	2.4255	2.5299	3.0297
6	6.1070	6.1070	6.1070	5.3465	5.4986	5.0892	4.7967
7	3.0189	4.4946	4.0027	3.7567	3.6092	3.5108	3.4405
8	5.9260	4.4612	4.9494	4.4556	4.1774	3.9907	3.8519
9	2.9520	2.9520	2.9448	2.9466	2.3573	2.9448	2.9459
10	2.9070	3.3451	3.3946	3.2732	3.2004	3.2492	3.3673

Figure 13: Simulation RMSFE

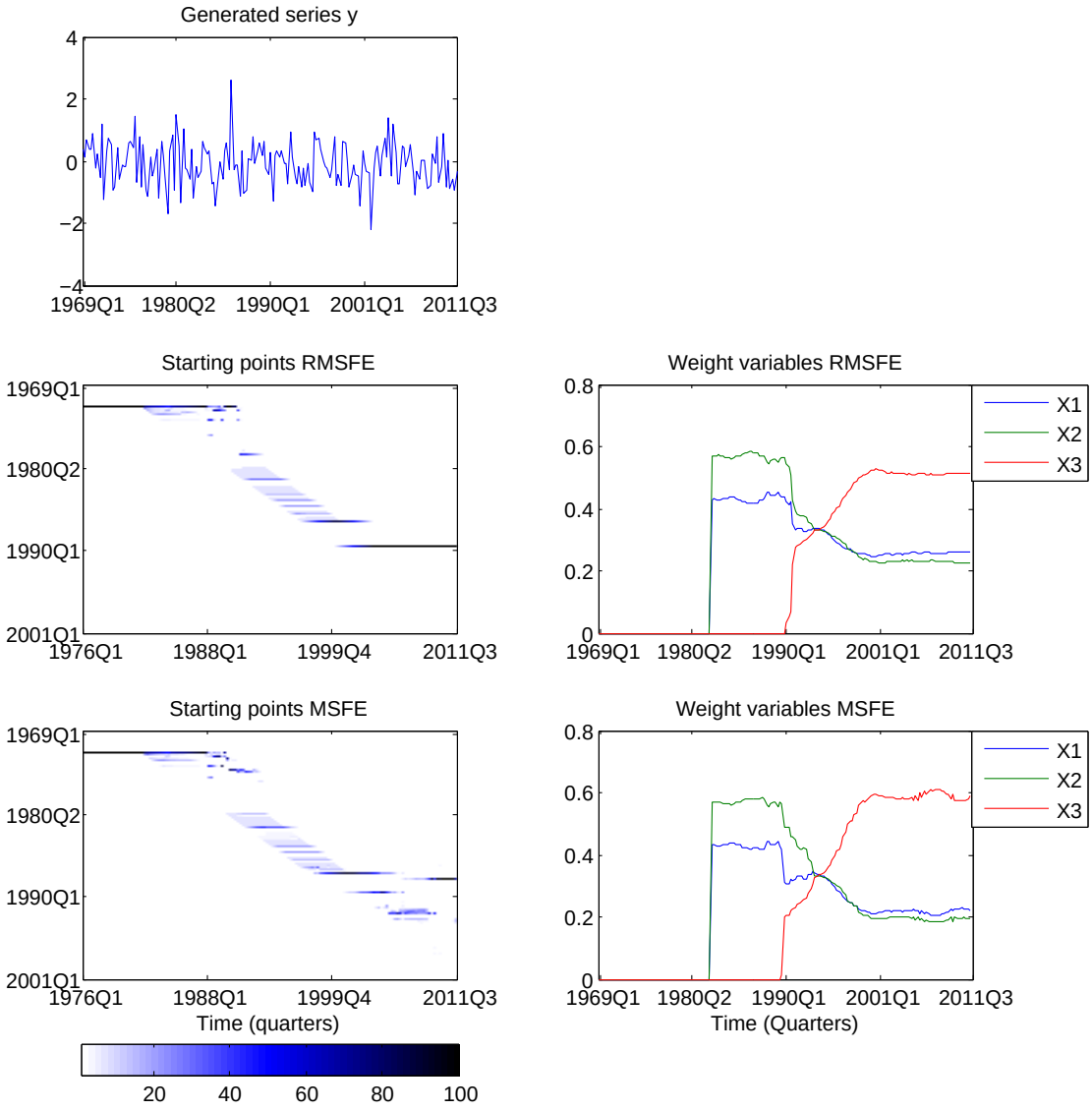


Figure 14: ***startSIMUL***: Overview shrinkage rate, selected variable, weight

RMSFE

TsimulRMSFE(1,71).resAAc(1,1).REST <172x7 double>

	1	2	3	4	5	6	7	8
88	0	0	0	0	0	0	0	
89	0	0	0	0	0	0	0	
90	0	0	0	0	0	0	0	
91	0	0	0	0	0	0	0	
92	0	3	1	2	0.5203	0.2628	0.2169	
93	0	3	1	2	0.5272	0.2586	0.2141	
94	0	3	1	2	0.5258	0.2594	0.2147	
95	0	3	1	2	0.5216	0.2518	0.2267	
96	0	3	1	2	0.5239	0.2510	0.2251	
97	0	3	1	2	0.5166	0.2641	0.2193	
98	0	3	1	2	0.5230	0.2609	0.2160	
99	0	3	1	2	0.5229	0.2683	0.2088	
100	0	3	1	2	0.5336	0.2728	0.1935	
101	0	3	1	2	0.5395	0.2656	0.1949	

MSFE

	1	2	3	4	5	6	7	8
88	0	0	0	0	0	0	0	
89	0	0	0	0	0	0	0	
90	0	0	0	0	0	0	0	
91	0	0	0	0	0	0	0	
92	0.3000	3	1	2	0.5899	0.2249	0.1851	
93	0.3000	3	1	2	0.5980	0.2198	0.1822	
94	0.3000	3	1	2	0.5964	0.2208	0.1828	
95	0.3000	3	1	2	0.5923	0.2147	0.1930	
96	0.3000	3	1	2	0.5950	0.2137	0.1913	
97	0.3000	3	1	2	0.5855	0.2269	0.1875	
98	0.4000	3	1	2	0.5560	0.2385	0.2054	
99	0.4000	3	1	2	0.5551	0.2444	0.2006	
100	0.4000	3	1	2	0.5641	0.2459	0.1900	
101	0.4000	3	1	2	0.5703	0.2393	0.1904	

Table 6: Overview functions used for generating results

File name	Input	Output	Short Description
indexBR	PGDP, IpredPGDP, NGDP,	Workspace with name:	Run index to run AA, TCOMB, and tabs/figs.
optpi	IpredNGDP,indPGDP	HF-PGDP-day-mo-year	indPGDP=1, PGDP is DV, else NGDP is DV.
Preparation			
prepare	PGDP, IpredPGDP	DV, Qtot, meanSPF, SPFr,aFE,sFE	Runs following functions:
realdata	PGDP (Real time realizations)	DV	DV, dependent variable, transformed forecasts.
dat	PGDP, Ipred (individual pred.)	[X, Ind]	Ind, forecasts of each forecaster, ordered by id.
qfind	Ind, DV	Qtot	Qtot, individual forecasts restructured.
ErInd	Qtot,DV	aFE,sFE,FE	Individual (squared/absolute) forecast errors.
SPFmed	Qtot,DV	meanSPF,numSPF, SPFr	Calculates mean SPF etc.
scoreRAE	Qtot,DV,aFE	W	Individual relative absolute error scores.
AA.a			
AAa	Qtot,DV,aFE,sFE ZQtot,ZDV,ZaFE,ZsFE, evalu	yAAa,S	Vars with first letter Z: prepare for score variable. yAAa struct. with Ma,...,Me. Runs:
optpi	Same as input AAa	S	Vary score hors/vars and track lengths in ranka
ranka	W, Qtot, ZW, scH,numlag, nc, scV, evalu	T	Rank experts; give score, rank, and forecasts
comboptmed	S	Ma,...,Me	Vary shrink. & score-hors in optmedShrink
optmedShrink	S,phi,maxc	MaW,...,MeW	Give shrunken inv. scores to top-ranked models.
ErrorM	Qtot,DV,Ma, ..., Me, evalu	SPFr,Eas,...,Ees	Give SE or RAE of all AA.a model forecasts.
TCOMB			
COMB	ypredtr,DV,j,SPFr,evalu	ONEypred, indSel, bench	Give shrunken inv. scores to opt. comb. of models.
TCOMB	DV,indM,Qtot,preds,evalu	y,startM,TresM, Tbench	Apply following functions.
TCOMBi	indM,DV,Qtot,preds,evalu,SPFr	startM	Compute model for different starting point.
indM=1: combAAa	start,DV,Qtot,yAAa,evalu	startM=num,yAAb,RES	AA.b: Sel yAAa, COMB score hor. with meanSPF.
indM=2: combAAb	start,yAAb1,yAAb2, yAAb3,DV,SPFr,evalu	startM=yAAc,resAAc	AA.c: COMB yAAb of each eval.funct.
indM=3 biasadjSPF	start,meanSPF,DV	startM=d,SICbias	SIC-Bias adjusted forecast.
TCOMBii	SPFr,DV,indM,startM,add,evalu	yii,REST,BENCH	COMB starting points
TCOMBiii	SPFr,DV,indM,startM,evalu	y,RESTTIME	Apply Tcombii for diff. sizes of moving window.

Note: When indented, the function is part of in the function above

Table 7: Overview functions used for displaying results

File name	Input	Output	Figure in paper/Short Description
tableRES	ypred,meanSPF,DV, SPFr, tt	A	Print results Table 1 in latex table format.
giveplots	DV,TresBIAS,TbenchBIAS,TresAAc1, TbenchAAc1,TresAAb1,TbenchAAb1, startAAb1,yAAa1,h,cmap,t,pl	time	pl=99: all plots in paper pl=n: particular figure in paper
retrieveMOD	num,Ma,disp,Rg	model, R	Retrieve model specs AA.a given number.
retieveNUM	M,Ma	r	Retrieve number 'r' given model AA.a specs.
permut	periods, opt	P	Handy for defining different permutations.
plotINTERVAL	DV,Qtot,yAAa1,yAAa2,yAAa3, yFINAL,meanSPF,indabs,h,t,pl		2.a: (pl=1) interval plot AA.a rel to AA. 2.b: (pl=2) histogram ranking AA in AA.a.
plotSTARTPOINT	Tres,h,indM,cmap	TTstart	3.a/4.a/8.a: DI starting points (SP).
plotselAAc	TresAAc, startAAc,h,cmap,pl	W,Mscore	3.b: DI of selected accuracy measures.
plotselAAb	TresAAb, startAAb,h,cmap	inde,Mscorehor	4.c: DI of selected score-horizons over time.
plotselAAa	TresAAb, startAAb, Ma,h	M,WE,WL,WS	4.b,d,e,f: DI of selected restrictions in Aa.a.
adhocSP	Tbench,DV,h,t,cmap,pl	Timage	8.c: (pl=1) image AE SP, 5.b: (pl=2) plot AE win, 5.b/8.d: (pl=3) plot AE SP
aposteriori	DV,Qtot,yAAa1, yAAa2, yAAa3, meanSPF,t		Table 2: *a posteriori* best and worst AA.a models.
adhocAAa	DV,yAAa1, yAAa2, yAAa3, meanSPF,h,t,cmap,pl	M,L,ST	6: (pl=1) Image AE of AA. T.4: (pl=3) S of AA.a 7: (pl=2) Response S to changing configurations.
plotselSIC	TresBIAS, startBIAS,3,cmap		8.e: DI of SIC choice Mean SPF vs Bias adj.
tableS	TbenchAAb1,TbenchAAc1, TbenchBIAS,DV		Table 3: S across windows and starting points.
simulxy	evalu,ty,tXstruc	tXstruc,ty,Tsimul, ysimul,RESTTIMEsimul	Simulation TCOMB.

Note: DI means distribution image.

References

Bache, I. W., Jore, A. S., Mitchell, J., and Vahey, S. (2012). Combining VAR and DSGE forecast densities. *Journal of Economic Dynamics & Control*, 35:1659–1670.

Capistrán, C. and Timmermann, A. (2009). Forecast Combination With Entry and Exit of Experts. *Journal of Business & Economic Statistics*, 27(4):428–440.

Gardner, E. (1985). Exponential Smoothing: The State of the Art. *Journal of Forecasting*, 4:1–28.

Hastie, T., Tibshirani, R., and Friedman, J. (2013 [2009]). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Heij, C., de Boer, P., Franses, P., Kloek, T., and van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics*. Oxford University Press.

Hendry, D. F. and Krolzig, H. M. (2005). The Properties of Automatic Gets Modelling. *The Economic Journal*, 115:32–61.

Hyndman, R. J. and Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 22:679–688.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. John Wiley & Sons, Inc.

Pesaran, M. H. and Timmermann, A. (2005). Real-Time Econometrics. *Economic Theory*, 21:212–231.

Pesaran, M. H. and Timmermann, A. (2007). Selection of Estimation Window in the Presence of Breaks. *Journal of Econometrics*, 137:134–161.

Stark, T. (2012). Survey of Professional Forecasters Documentation. *Federal Reserve Bank of Philadelphia*, pages 1–41.

Stock, J. H. and Watson, M. W. (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting*, 23:405–430.

Zarnowitz, V. and Braun, P. (1992). Twenty Years of the NBER-ASA Quarterly Economic Outlook Surveys. *National Bureau of Economic Research*, 3965.